COMPARATIVE ANALYSES IDENTIFY ADAPTIVE GENETIC VARIATION IN CROPS AND CROP WILD RELATIVES

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

ZHOU FANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISOR: PETER L. MORRELL

DECEMBER 2013

ACKNOWLEDGEMENTS

I thank my advisor Peter L. Morrell. Peter has provided me with several exciting projects, which have gradually brought me into the field of population genetics and evolutionary biology and have helped me develop expertise in this area. Without Peter's guidance, I would not be at this stage of my scientific career. Peter is always very patient when explaining complex concepts. He also helps me improve my writing and presentation skills and always provides valuable and prompt comments on every manuscript and presentation. I will apply the knowledge I have learned from him to solve new problems and make my own discoveries along the way.

I thank my committee members Gary Muehlbauer, Molly McCue, David Moeller and Peter Tiffin for their help during each stage of my PhD and for their support and trust that I could develop the skills to complete this PhD. I thank them for all the valuable comments and suggestions they have provided on my projects and their advice and input in every committee meeting.

I thank the Directors of Graduate Studies of the Plant Biological Sciences graduate program during my PhD study: Jane Glazebrook, Gary Muehlbauer, George Weiblen, and Cindy Tong. I thank them for helping me start my PhD, writing letters for me, and making suggestions on my application materials when I applied for fellowships. I also thank the Plant Biological Sciences graduate program coordinator Gail Kalli. Gail has helped me in many different ways before and during my PhD study. Gail is always very

patient and helpful. Gail provides prompt response whenever I have questions and always reminds me if I forget anything.

I thank everyone in the Morrell Lab, especially Ana Gonzales and Thomas Kono. Ana and Tom are extremely understanding and caring colleagues and friends. They are always actively involved in discussion of my projects, providing comments and edits on my manuscripts and presentations. Thank you to all other members in the Morrell Lab: Diego Coelho, Kevin Volz, Amber Eule-Nashoba, Beau Miller, Kiran Seth, Chaochih Liu, and Ashley Rozmarin, for help on my projects. I also thank everyone in the Muehlbauer and Stupar Labs for the discussions, and for the dynamic and friendly work environment they have provided me these four years.

I thank Michael Clegg from University of California, Irvine and Jeffrey Ross-Ibarra from University of California, Davis for their guidance on my projects. I appreciate their comments and edits on my papers and helpful discussions on my projects. Their knowledge, attitude and passion have influenced me gradually during my PhD and brought unexpected positive impacts on my PhD study and my career as a scientist.

I appreciate funding from the Plant Biological Sciences graduate program for supporting my thesis research and for defraying the cost of attending international conferences and workshops. I am very grateful for the Doctoral Dissertation Fellowship from the University of Minnesota Graduate School. This support was critical as I have finished my research and prepared my dissertation. I also wish to thank Syngenta for providing me with two internship positions and thus the opportunity to apply what I have learned during my PhD to solve real world problems.

Chapter 2: I am the first author on this project. Tanja Pyhäjärvi, Allison L. Weber, R. Kelly Dawe, Jeffrey C. Glaubitz, José de Jesus Sánchez González, Claudia Ross-Ibarra, John Doebley, Peter L. Morrell, and Jeffrey Ross-Ibarra are co-authors on this paper. T.P., P.L.M., J.R-I and I were the primary contributors to the published work in the project; J.C.G. and J.J.S.G. collected the original samples and genotype data; A.L.W. and J.D. contributed genetic maps; C.R-I. and R.K.D. performed the cytology study; T.P. did the association analysis; I performed almost all other research and analyzed data; P.L.M., J.R-I. and I wrote the paper.

We thank Kate Hodges for help with cytological scoring, Kevin Volz for help with geographic information systems data and for producing Figure S2.3, and Graham Coop, Loren Rieseberg and two anonymous reviewers for helpful comments on earlier versions of this manuscript. This work was carried out using computing resources at the University of Minnesota Supercomputing Institute. The authors wish to acknowledge funding from the Academy of Finland to T.P., start-up funds from the University of Minnesota Department of Agronomy and Plant Genetics to P.L.M., and from the National Science Foundation (NSF) (DBI-0820619 to J.D. and NSF IOS-0922703 to R.K.D. and J.R-I.).

Chapter 3: I am the first author on this project. Amber Eule-Nashoba, Carol Powers, Thomas Y. Kono, Shohei Takuno, Peter L. Morrell, and Kevin P. Smith are co-authors on this paper. P.L.M., K.P.S. and I contributed most part on this project; I performed all research and analyzed data; A.E-N. made Figure 3.1; C.P. and K.P.S. contributed the data on the breeding population; T.Y.K., S.T. and P.L.M. provided technical help on

simulations; P.L.M., K.P.S. and I wrote the paper.

The authors thank Ed Scheifelbein, Karen Beaubien, Shiaoman Chao for SNP genotyping. Ana Gonzales, Matthew Hufford, Sofiane Mezmouk, Mohsen Mohammadi, Tanja Pyhäjärvi, Jeffrey Ross-Ibarra and two anonymous reviewers provided discussion and helpful comments on an earlier version of the manuscript. This work was carried out in part using hardware and software provided by the University of Minnesota Supercomputing Institute. We acknowledge funding from the US Department of Agriculture National Institute for Food and Agriculture USDA NIFA, 2006-55606-16722 for support to K.P.S., USDA NIFA 2011-68002-30029 for support to K.P.S. and P.L.M., USDA NIFA 2011-38420-20068 for support to P.L.M., and the University of Minnesota Faculty Grant-in-Aid of Research to P.L.M.

Chapter 4: I am the primary researcher and author on this project. Ana M. Gonzales, Michael T. Clegg, Kevin P. Smith, Gary J. Muehlbauer, Brian J. Steffenson, and Peter L. Morrell are co-authors on this paper. P.L.M. and I were the primary contributors to the project; I performed all research, analyzed data and wrote the paper; B.J.S. contributed the data; A.M.G., M.T.C., K.P.S., G.J.M., B.J.S. and P.L.M. participated in the discussion; A.M.G, M.T.C., G.J.M., B.J.S. and P.L.M. provided comments and edits on an earlier version of the manuscript.

This work was carried out using computing resources at the University of Minnesota Supercomputing Institute. We acknowledge funding from the US Department of Agriculture National Institute for Food and Agriculture USDA NIFA 2011- 68002-30029 to P.L.M. and University of Minnesota Doctoral Dissertation Fellowship to Z.F. The

authors thank Thomas Kono for assistance with determination of SNP ancestral state and helpful discussion; Allison Haaning, Tanja Pyhäjärvi, Jeremy Yoder for comments on an earlier version of the manuscript.

ABSTRACT

Comparative population genetic analyses provide a means of identifying adaptive genetic variation. In this dissertation, I apply population genetic approaches to identify putatively adaptive variants in the genomes of crops and crop wild relatives. These approaches have the potential to identify genetic variants that are under selection and thus potentially contributing to local adaptation. As a background to the dissertation, I present in Chapter 1 the state of research in this field at the time I started my PhD and give a brief introduction to the projects described in this dissertation. In Chapter 2, I report a ~50-Mb chromosomal inversion in the wild ancestor of maize – teosinte (*Zea mays* ssp. *parviglumis*) and characterized its distribution and abundance in natural populations using population genetic approaches. This is also the first study in plants to apply population genetic approaches to identify chromosomal structural variation. In Chapter 3, I used a population genetic approach to identify genomic regions that contain adaptive mutations resistant to *Fusarium* head blight in a barley experimental breeding population. The successful application of comparative population genetic approaches in this study suggests this approach can also be used to identify genomic regions that are under selection in other breeding populations. In Chapter 4, I studied the geographic differentiation in wild barley (*Hordeum vulgare* ssp. *spontaneum*). I found two genomic regions contribute disproportionately to the population structure in wild barley. These same regions, with reduced evidence of recombination, are strongly associated with environmental variables. Population genetic evidence and previous cytological and

genetic studies suggest these two genomic regions may be chromosomal structural rearrangements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Crop improvement depends on our ability to identify favorable genetic variants and to combine them in modern breeding programs. Mutation and recombination are two sources of genetic variations. Mutation creates new genetic variants. Recombination rearranges those variants into new combinations, so it has the potential to combine favorable mutations onto the same chromosome to improve overall fitness (Hill and Robertson, 1966). The contribution of mutation and recombination to genetic diversity differs dramatically across the genome (Gore *et al.*, 2009; Mayer *et al.*, 2012), in different subspecies or different populations of the same species (Ross-Ibarra *et al.*, 2009; Fang *et al.*, 2013a) and in different species (Morrell *et al.*, 2006; Gonzales *et al.*, 2012).

Under a standard neutral model, levels of linkage disequilibrium (LD) are high when recombination is limited in a chromosomal segment. The extent and distribution of LD in the genome can contribute to the identification of adaptive genetic variants. Association (or LD) mapping permits the association of phenotypes and genotypes with higher genetic resolution than previous methods and has found broad applications in the search for disease-associated human genes and identification of loci controlling important crop traits (eg., Mackay *et al.*, 2009; Cockram *et al.*, 2010; Huang *et al.*, 2010; Kump *et al.*, 2011). Among the most crucial factors for association mapping is the level of LD between genetic markers and causative mutations that contribute to observed phenotypes (Long and Langley, 1999). One limitation of these approaches is that the phenotypes need to be identified and accurately measured (Ross-Ibarra *et al.*, 2007).

In contrast to these top-down approaches, bottom-up approaches do not need phenotype information *a priori*, but rather use population genetic analyses to identify

potentially adaptive genetic variants and then connect these genetic variants to phenotypes (Ross-Ibarra *et al.*, 2007). The fundamental framework derives from the observation that demographic effects on populations tend to impact all loci while selection acts on individual loci (Cavalli-Sforza, 1966). This observation leads to explicit test for loci that differ in frequency among populations (Lewontin and Krakauer, 1973). For bi-allelic SNPs, F_{ST} measures divergence in minor allele frequency, with larger differences between two populations creating higher F_{ST} (Weir and Cockerham, 1984). Loci with outlier F_{ST} values suggest they are targets of selection or more likely, linked to a target of selection through genetic hitchhiking. However, this approach is limited by the high degree of variance in expected allele frequency between populations (Nei and Maruyama, 1975).

Comparison of allele frequency differences with F_{ST} requires the prior identification of populations for comparison, a problem directly addressed by environmental association approaches, such as Bayenv (Coop *et al.*, 2010; Günther and Coop, 2013) and Spatial Ancestry Analysis (SPA) (Yang *et al.*, 2012). Differences in allele frequency at adaptive genetic variants are usually driven by environmental variables and correlated selection pressures. Environmental association attempts to identify genetic variants that differ in frequency among populations adapted to different environmental conditions. Bayenv is based on localized small clusters to correct for population structure in environmental association (Coop *et al.*, 2010). SPA is used to model how the allele frequency of each SNP changes as a function of the location of the individual in geographic space. The SPA method is particularly sensitive to SNPs that have steep

geographic gradients in allele frequency. In contrast to Bayenv and F_{ST}, the SPA method is scored based on individual accessions rather than populations (Yang *et al.*, 2012).

Population genetic approaches can help identify adaptive genetic variations, but where are these adaptive genetic variations most likely to be located in the genome? Both theory and empirical observations suggest that chromosomal inversions are especially likely to harbor adaptive variations (Kirkpatrick and Barton, 2006; Hoffmann and Rieseberg, 2008). Inversions are particularly disruptive because a single crossover in an inversion region produces inviable and unbalanced gametes (Burnham, 1962). In the absence of a selective advantage, inversions are quickly lost from populations by purifying selection owing to the fitness cost of unbalanced gametes and the potential for inversions to harbor deleterious variants (Kirkpatrick and Barton, 2006; Guerrero *et al.*, 2012). Despite the fitness effects associated with disruption of meiotic pairing, inversions are strongly favored when they serve to protect locally adapted mutations from gene flow just by capturing alleles at two or more loci that improve local adaptation (Kirkpatrick and Barton, 2006). Due to the potential of local adaptation, many inversions display strong clinal variation. This relationship between geographic distribution and inversion is often due to climate difference or water availability (Levitan, 2001; Umina et al., 2005; Balanyá *et al*., 2006; Lowry and Willis, 2010; Ayala *et al.*, 2011).

While cytological studies have characterized many inversions in plants, including teosinte and wild barley (eg., Ting, 1976; Konishi and Linde-Laursen, 1988), population genetic approaches provide a much more rapid means of detecting putative inversions. Population genetic comparisons also offer the only tractable means of characterizing

inversion frequency and distribution based on the large amount of samples from natural populations (Fang *et al.*, 2012). Inversions inhibit recombination between the inverted and ancestral chromosomal arrangements in heterozygotes, thus the coalescence times for these two arrangements increases (Guerrero *et al.*, 2012). Therefore, population genetic evidence of inversion includes extended LD, highly divergent haplotypes, and increased sequence differentiation (Munte *et al.*, 2005; Huynh *et al.*, 2011; Cheng *et al.*, 2012).

There are still many questions about the nature of local adaptation and the potential role of inversions, that remain poorly understood (Kirkpatrick and Kern, 2012). In particular, the abundance and adaptive contribution of inversion polymorphisms in natural populations has only begun to be addressed. The second chapter of this dissertation is an article published in Genetics (Fang *et al.*, 2012). I am the first author of this paper; the contributions of coauthors are detailed in the acknowledgements section. In brief, we discovered a novel ~50-Mb inversion (*Inv1n*) in teosinte, the wild ancestor of maize. The population genetic data set consisted of ~1000 SNPs in > 1000 individuals sampled from 33 populations. This inversion is the largest identified in plants. The SNP data demonstrate extended LD, highly divergent haplotypes, and high differentiation in resequencing data. We present evidence that the rearrangement is relatively ancient and has persisted in teosinte populations for sufficient time to permit its widespread occurrence in all natural populations investigated. Our data further suggest the rearrangement is adaptive, as the rearrangement frequency shows a strong correlation with altitude and is associated with multiple environmental variables and phenotypes. We speculate that the variation in temperature or precipitation, that correlated with altitude,

contribute to the selective pressure. We found this inversion is not present in maize by studying the same set of genotypes in 1,573 maize samples, thus the variants in *Inv1n* could potentially improve maize adaptation to lower rainfall or higher heat regimes.

The third chapter of my dissertation is an article published in G3: Genes | Genomes | Genetics (Fang *et al.*, 2013b). I am the first author of this paper; the contributions of coauthors are detailed in the acknowledgements section. We used population genetic approaches to identify genomic regions showing evidence of change in allele frequency in response to selection for lines with resistance to *Fusarium* head blight. The population was subject to introgression from 13 disease resistant donors. Using comparative analyses between the original population and the population subject to introgression, we have identified genomic regions and likely donors most responsible for increased disease resistance based on differentiation in allele frequency between populations, extended LD, and identification of genomic regions with increased identity-by-state to donors. I also used coalescent simulations to show that these genomic patterns are not likely to arise due to demography alone without imposing selective pressure. These genomic regions overlap regions previously identified from quantitative trait locus (QTL) mapping and association mapping.

In the fourth chapter, I studied geographic structure and genetic differentiation in wild barley using 3,072 genetic variants in a sample of 318 accessions. I am the first author of this manuscript; the contributions of coauthors are detailed in the acknowledgements section. We found that wild barley accessions are differentiated into two major populations, divided west and east of the Zagros Mountains and the genetic

differentiation is contributed primarily by two genomic regions, one on linkage group 2H and the other on 5H. These two genomic regions contain markers with the highest informativeness for assignment, the largest allele frequency differences between the two primary populations, and the largest gradients in allele frequency based on isolation-by-distance. Previous cytological and genetic studies suggest there are chromosomal translocation or inversion in these two genomic regions (Ramage and Suneson, 1961; Konishi and Linde-Laursen, 1988). The putative chromosomal structural variation on 2H is associated with both temperature and precipitation variables while the putative inversion on 5H is associated with precipitation variables. Based on the current marker density, most adaptive genetic variations identified are captured by these two chromosomal structural variations, which are the regions we are most likely to detect, because these regions have relatively high LD and selection for local adaptation preserves these regions.

CHAPTER 2

MEGABASE-SCALE INVERSION POLYMORPHISM IN THE WILD ANCESTOR OF MAIZE

Chromosomal inversions are thought to play a special role in local adaptation, through dramatic suppression of recombination, which favors the maintenance of locally adapted alleles. However, relatively few inversions have been characterized in population genomic data. Based on single nucleotide polymorphism (SNP) genotyping across a large panel of *Zea mays,* we have identified an ~50-Mb region on the short arm of chromosome 1 where patterns of polymorphism are highly consistent with a polymorphic paracentric inversion that captures more than 700 genes. Comparison to other taxa in *Zea* and *Tripsacum* suggests that the derived, inverted state is present only in the wild *Zea mays* subspecies *parviglumis* and *mexicana,* and is completely absent in domesticated maize. Patterns of polymorphism suggest that the inversion is ancient, and geographically widespread in *parviglumis*. Cytological screens find little evidence for inversion loops, suggesting that inversion heterozygotes may suffer few cross-over induced fitness consequences. The inversion polymorphism shows evidence of adaptive evolution, including a strong altitudinal cline, a statistical association with environmental variables and phenotypic traits, and a skewed haplotype frequency spectrum for inverted alleles.

2.1 Introduction

The evolutionary role of chromosomal inversions has been studied in a wide array of organisms, from insects (Ayala *et al*., 2011; Stevison *et al*., 2011) to birds (Huynh *et al*., 2011) and plants (Hoffmann and Rieseberg, 2008; Lowry and Willis, 2010). Examination of inversion polymorphism was fundamental to the early study of selection and adaptive diversity, as well as the basis for understanding the maintenance of neutral polymorphism within populations (Dobzhansky, 1950; Hoffmann *et al*., 2004). Homologous pairing of an inverted and non-inverted chromosome in heterozygotes leads to the formation of an inversion loop, and crossing over in an inversion loop can cause the formation of a dicentric chromosome and an acentric fragment at meiosis I, resulting in terminal deletions of the affected chromosome and gamete death at frequencies that correlate with the size of the inversion (Burnham, 1962). Because of the difficulty of homologous pairing and the deleterious effects of homologous crossing over in inversions, inversions are typically observed to disrupt recombination in heterozygous individuals, leading to measurable effects on nucleotide sequence polymorphism, including the generation of extended LD. Inversion induced LD has been reported in a variety of organisms, including humans (Bansal *et al*., 2007), *Drosophila subobscura* (Munte *et al*., 2005) and several other species (reviewed in Hoffmann and Rieseberg, 2008). Strong differentiation between chromosomal arrangements (as measured by F_{ST}) has also been used as evidence of inversions in *Drosophila* (Andolfatto *et al*., 1999; Depaulis *et al*., 1999; Nóbrega *et al*., 2008).

A variety of circumstances can favor the maintenance or spread of an inversion polymorphism. The inversion may be selected for if the structural rearrangement itself has fitness consequences (Castermans *et al*., 2007). Natural selection can also favor the spread of an inversion if it contains locally adapted alleles, because inversions can suppress recombination and thus protect adaptive alleles from gene flow (Kirkpatrick and Barton, 2006; Machado *et al*., 2007). Some inversion polymorphisms display strong patterns of geographic structure, consistent with local adaptation to ecological factors such as temperature regimes or water availability (White *et al*., 2009; Lowry and Willis, 2010; Ayala *et al*., 2011). For example, strong differentiation among ecological zones was observed for an inversion in the mosquito *Anopheles funestus* (Ayala *et al*., 2011). In the yellow monkeyflower, *Mimulus guttatus* (Lowry and Willis, 2010), an inversion is involved in local adaptation to Mediterranean habitats through several morphological and phenological traits, while the standard arrangement appears in a perennial ecotype from habitats with high year-round soil moisture. In addition to selection, inversion polymorphisms without strong deleterious effects may also increase in frequency through genetic drift and migration, potentially resulting in fixation in small populations (Bengtsson and Bodmer, 1976; Lande, 1984).

Here we examine the population-level diversity of a newly discovered 50 Mb inversion found in the wild subspecies of *Zea mays* (known collectively as teosinte). Inversions in both wild and domesticated *Zea mays* and related taxa have been reported previously (McClintock, 1931; Morgan, 1950; Ting, 1965; Ting, 1967; Kato Y., 1975; Ting, 1976), but these were detected cytologically and little is known about their

evolution in natural populations. We examine genome-wide patterns of LD in *Zea mays* using 941 SNP markers genotyped in a diverse sample of 2782 individuals, including representatives of three *Zea mays* subspecies: domesticated maize (*Zea mays* ssp. *mays*), its wild progenitor *Zea mays* ssp. *parviglumis*, and the weedy taxon *Zea mays* ssp. *mexicana* (hereafter *mays*, *parviglumis*, and *mexicana*, respectively). A region spanning ~50 Mb on the short arm of chromosome 1 in *parviglumis* and *mexicana* demonstrates the highest level of LD in the genome and coincides with a region of high differentiation between *mays* and *parviglumis* reported by Hufford *et al*. (2012). Comparison to other taxa in *Zea* and *Tripsacum* suggests that the inverted arrangement is derived. The inverted arrangement is present at population frequencies up to 90% in *parviglumis,* but completely absent in domesticated maize. We present evidence that the inversion is relatively ancient and has persisted in teosinte populations for sufficient time to permit its widespread occurrence in all 33 natural populations of subspecies *parviglumis* investigated. Our data further suggest the inversion may be adaptive, as the inverted arrangement shows a strong altitudinal cline, is associated with multiple environmental and phenotypic traits, and the haplotype frequency spectrum of inverted alleles appears inconsistent with neutral evolution.

2.2 Materials and Methods

2.2.1 Plant materials and genetic data

Plant materials (available at http://www.genetics.org/) included accessions of all four subspecies of *Zea mays* (1573 ssp. *mays*, 975 ssp. *parviglumis*, 161 ssp. *mexicana*, and 10 ssp. *huehuetenangensis*), as well as *Z. luxurians* (17), *Z. diploperennis* (15), and *Z. perennis* (9). The panel also included 22 *Tripsacum* accessions used as outgroups. The 975 *parviglumis* accessions include 33 populations with at least 10 individuals in each population. The *mays* samples include 1283 accessions representing approximately 250 traditional open-pollinated landraces (including 27 inbred landraces), and 290 modern inbred lines.

Genotyping for the genome-wide set of 959 SNPs followed previously described methods (Weber *et al*., 2007; van Heerwaarden *et al*., 2010). The SNP discovery panel consisted of 14 *mays* inbred lines and 16 teosinte partial inbreds (Wright *et al*., 2005; Weber *et al*., 2007). We excluded accessions (seven *parviglumis*, one *mexicana* and 20 *mays*) and loci (three SNPs) with more than 15% missing data. We also removed SNPs where BLAST searches of context sequence identified multiple locations in the *mays* reference genome (Release 5a.59) (Schnable *et al*., 2009). This resulted in a final panel of 941 SNPs from 542 mapped genes. SNP genotypes and their contextual sequences are available at http://www.panzea.org and http://www.rilab.org. A subset of these data have been published elsewhere, including 706 SNPs from 584 *parviglumis* accessions in

Weber *et al*. (2007), 123 SNPs from 817 *parviglumis* in Weber *et al*. (2008) and 468 SNPs from 1127 *mays,* 100 *parviglumis,* and 96 *mexicana* in van Heerwaarden *et al*. (2011).

The majority of the accessions represent open-pollinated populations that are highly heterozygous, resulting in genotypic data of unknown phase. We computationally phased each of the three subspecies separately using the software fastPHASE (Scheet and Stephens, 2006) with 20 random starts and 25 iterations of the Expectation-Maximization algorithm. For each subspecies, inbred lines were used as training data.

We made use of published Sanger resequencing data from Wright *et al*. (2005). Wright and coauthors sequenced PCR products from teosinte accessions which had been inbred for two generations. Within the inversion, two individuals with a sequence from one arrangement at one locus and the alternate arrangement at a different locus were deemed heterozygous and removed from the analysis. We limited our analyses to reliably aligned loci on chromosome 1 with $n \geq 10$ total sequences and $n \geq 3$ of each arrangement. In total, we analyzed 95 loci, including seven inside the inversion (data available at http://www.genetics.org/ and www.panzea.org). We used the top BLAST hit from the *Sorghum* genome (Paterson *et al*., 2009) as an outgroup for each locus. Loci were annotated by BLAST comparison to the *mays* reference genome (Release 5a.59) (Schnable *et al*., 2009).

2.2.2 Data analysis and divergence time

Summary statistics of the SNP genotyping data and the resequencing data were

calculated using the analysis package of the libsequence library (Thornton, 2003). Although the absolute values of some summary statistics of the SNP genotyping data may be affected by ascertainment bias (Clark *et al*., 2005a), the relative values of these statistics are expected to be more robust.

We tested for Hardy-Weinberg equilibrium by treating the inversion as a single biallelic locus. LD (as measured by r^2) was calculated for SNPs with a minor allele frequency (MAF) >5% in R (R Development Core Team, 2011) and LDheatmap package (Shin *et al*., 2006) was used to plot LD. To assign individuals to haplotype clusters at the inversion, we used the genetic assignment software STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003). We estimated haplotype clusters for values of K ranging from 1 to 5. For each value of K, we used 10 replicate runs of the admixture model, with a burn-in of 100,000 iterations and a run length of 100,000 steps. To compare differentiation inside and outside of the inversion, we divided the sample of *parviglumis* into the two clusters identified by STRUCTURE and calculate F_{ST} (Weir and Cockerham, 1984) between these two groups along chromosome 1.

Genetic (Manhattan) distance among inbred *parviglumis* lines was estimated in the software TASSEL (Bradbury *et al*., 2007) and calculated separately for SNPs inside and outside of the inversion. We used a Fitch-Margoliash least squares approach (Fitch and Margoliash, 1967) as implemented in the software package PHYLIP (Felsenstein, 1993) to estimate a dendrogram for all taxa using the 17 SNPs inside the inversion.

We applied two common tests of neutrality (Hudson *et al*., 1987; McDonald and Kreitman, 1991) to the Sanger resequencing data. For the McDonald-Kreitman (MK) test

we used the seven resequencing loci inside the inversion and compared polymorphism at the inverted arrangement to divergence between the inverted arrangement and a *Sorghum* outgroup. For the Hudson-Kreitman-Aguade (HKA) test, we used 74 loci with ancestral information. HKA tests were performed both for the combined set of sequences as well as for sequences from each of the chromosomal arrangements separately. Because loci within the inversion are unlikely to be independent, we summed polymorphism and divergence data across loci within the inversion. We used the maximum likelihood approach of Wright and Charlesworth (2004), running 100,000 Markov chain Monte Carlo (MCMC) iterations, and a starting *parviglumis*-*Sorghum* divergence of 60N generations.

We estimated divergence time between the arrangements using sequences at seven loci from the two observed haplotype groups inside the inversion. We treated samples of the two chromosomal arrangements as distinct populations, and estimated divergence time under an isolation with migration model as implemented in the software MIMAR (Becquet and Przeworski, 2007). We set the inheritance and the mutation rate variation scalars both to 1, and the recombination inheritance and rate variation scalar to $(Z_{ri}\text{-}1)/(Z_i\text{-}1)$, where Z_{ri} is the initial length of locus *i* and Z_i corresponds to the number of base pairs in locus *i* after filtering out indels and missing data. The mutation rate per generation per base pair was assumed constant across loci and set to 3×10^{-8} (Clark *et al*., 2005b). The population mutation rate per base pair, divergence time, and the natural logarithm of the population migration parameter were sampled from the uniform distributions U(0, 0.08), U(0, 10^6), and U(-2, 1), respectively. The exponential growth

parameter was set to six (other values did not change results considerably). We ran the Markov chain for 5,000 burn-in steps followed by 10,000 steps for parameter estimation, repeating our analysis with two independent seeds. Three hundred genealogies were generated per locus for each step of the MCMC. We inferred that convergence was reached when the posterior distributions of both runs were very similar; results reported are the average of both runs.

2.2.3 Association analyses

We used a Bayesian approach (Coop *et al*., 2010) to test for associations between the inversion and 22 geographical (altitude, latitude, and longitude) and bioclimatic variables (worldclim.org) (Hijmans *et al*., 2005) in the 33 *parviglumis* populations with ≥10 samples (Table S2.1). The analysis explicitly accounts for population structure using a covariance matrix of allele frequencies estimated by 50,000 MCMC steps using all SNPs. We assessed association genome-wide using all SNPs, and using a single-marker test treating the inversion as a single locus. In each case, five separate runs with 50,000 iterations were performed to control for differences among MCMC runs.

To examine whether the inverted arrangement was associated with phenotypic variation, we used phenotype data from Weber *et al*. (2008). Both phenotype and genotype data were available for 811 individuals. A kinship matrix was estimated from all 941 SNPs using the options "all" and "additive" in the EMMA R package (Kang *et al*., 2008). Both genome-wide association and single-marker association (treating the inversion as a single locus) were performed for each of 37 phenotypes. Associations were

tested using a mixed linear model as implemented in the software TASSEL (Bradbury *et al*., 2007). The R package qvalue was used to estimate the false discovery rate (FDR) and identify SNPs that were significant at an FDR of 5% (Storey and Tibshirani, 2003).

2.2.4 Cytology

To assess the potential cytological impacts of the inversion, we screened 174 meiocytes from immature tassels of six *parviglumis/mays* F1 progeny resulting from the cross of two inbred *parviglumis* each with a single *mays* inbred line. Meiocytes were collected and staged following Li *et al*. (2010) and the chromosomes stained with either 1% aceto-orcein or DAPI. Recombination within the inversion was scored as previously described, noting chromosome bridges and acentric fragments at anaphase I (Dawe and Cande, 1996).

2.3 Results

We examined the level of LD in each of the three subspecies of *Zea mays* with a genome-wide set of 941 SNPs from 2782 samples. Using computationally phased genotypic data, we searched for pairs of markers in high LD ($r^2 > 0.6$) and separated by >1 Mb. Our scan identified two such regions, an ~50-Mb region on chromosome 1 and a ~15 Mb span of chromosome 8. Because the region on chromosome 8 is near a likely assembly error in the reference genome (J. Glaubitz unpublished data), we focused our analysis on chromosome 1. The region of high LD on chromosome 1 in our data corresponds closely to the region Mb 65-115 on the physical map of the reference *mays* genome (B73 RefGen v2, release 5a.59, 2010-11) recently reported by Hufford *et al*. (2012) as a putative inversion. Our data reveal high LD (mean $r^2 = 0.24$) among the 17 SNPs from Mb 65.09 to 106.16 (Figure 2.1), compared to a genome-wide average of 0.004. Gametic disequilibrium, as estimated from unphased SNP genotyping data, also demonstrates this excess of LD (data not shown). Finally, high levels of LD are also evident in genotypic data from a panel of 13 individuals of *parviglumis* genotyped using the 55,000 SNPs on the MaizeSNP50 Illumina Infinium Assay (Hufford *et al*., 2012), suggesting that the LD observed is not an artifact of the genotyping platform used.

2.3.1 The extended region of high LD on chromosome 1 is a putative inversion

Because *mays* and the teosintes are outcrossing taxa with large effective population

sizes, LD in the genome generally declines rapidly with distance ($r^2 < 0.1$ within 1500 bp in domesticated *mays*) (Remington *et al*., 2001). The region of high LD is distinct from both the centromere (Wolfgruber *et al*., 2009) and known heterochromatic knobs (Buckler *et al*., 1999) and exhibits relatively low recombination (Figure 2.1). An ~50-Mb span of high LD is unexpected, and while *parviglumis* and *mexicana* show evidence of high LD in this chromosomal region, levels of LD in our large sample of domesticated *mays* are similar to genome wide averages (Figure 2.1). Other wild taxa also do not show an excess of LD on the short arm of chromosome 1, although our power to measure LD in these samples is likely hampered by smaller sample size and SNP ascertainment bias. Finally, a recent genetic map from a BC2S3 population derived from a cross between a *mays* line and a *parviglumis* line with the putatively inverted arrangement shows no crossovers inside the ~50-Mb span in the 881 progeny genotyped, consistent with the putative inversion suppressing recombination in heterozygotes (L. Shannon and J. Doebley unpublished data). Though final validation will require demonstrating differential marker order in the progeny of self-fertilized individuals homozygous for alternate arrangements (Mano *et al*., 2012), we view these multiple lines of evidence as a strong case that recombination is suppressed due to an inversion in this region, henceforth identified as *Inv1n.*

To test for evidence of pairing and recombination within the large *Inv1n* region, we examined male meiocytes from six F1 plants derived from two crosses between *mays* and an inbred *parviglumis* line containing *Inv1n*. Both hybrids revealed a low frequency of dicentric bridge formation at ~4% (7/167), but no acentric fragments were observed

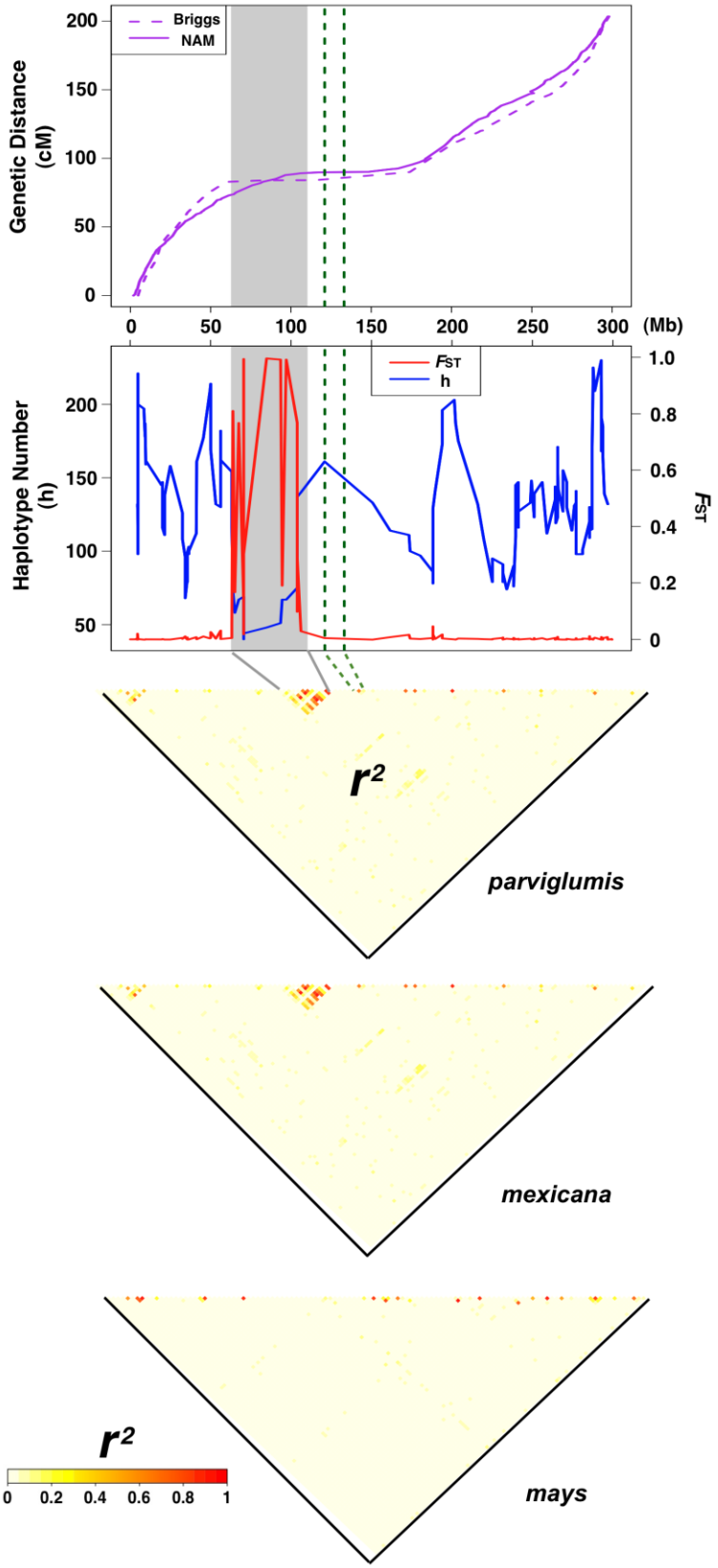

Briggs
NAM
Genetic Distance (cM)
200
150
100
50
0
0
50
100
150
200
250
300
(Mb)
FST
h
Haplotype Number (h)
200
150
100
50
1.0
0.8
0.6
0.4
0.2
0
FST
r2
parviglumis
mexicana
mays
r2
0
0.2
0.4
0.6
0.8
1

Figure 2.1 Population genetic evidence for the *Inv1n* inversion.
The top panel shows cumulative genetic distance by physical position along chromosome 1. The dashed curve is based on the teosinte-maize backcross map of Briggs *et al*. (2007) and the solid curve from the maize nested association mapping (NAM) population (Yu *et al*., 2008). The bottom panel shows haplotype number (blue curve) and F_{ST} between the inverted and standard arrangements (red curve). The number of haplotypes present across chromosome 1 was calculated in overlapping 10-SNP windows with 1 SNP increments. The inverted region is marked in grey, and the centromere by green dashed lines. Below the panels, LD (r^2) is plotted across the chromosome for *parviglumis*, *mexicana*, and *mays*.

(Table S2.2). Although such bridges were rare, an anaphase I bridge in a plant heterozygous for *Inv1n*was observed (Figure S2.1). In addition, we observed no obvious reduction in pollen viability or seed set in a total of five F1 plants (data not shown).

2.3.2 Haplotype variation and divergence time

STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003) analysis of SNPs on all 1936 *parviglumis* chromosomes inside *Inv1n* shows the highest likelihood for $K = 2$ clusters, a pattern not seen from the full set of genome-wide SNPs (data not shown). These groups are hereafter referred to as *Inv1n-I* and *Inv1n-S* for the inverted and standard arrangements, respectively (Figure 2.2). Recombination among loci within a chromosomal arrangement should be unaffected, and levels of LD within *Inv1n-I* (mean $r^2 = 0.11$) and *Inv1n-S* arrangements (mean $r^2 = 0.07$) are indeed low and similar to background levels (Figure S2.2). Average F_{ST} between chromosomes with alternate arrangements is notably higher inside the *Inv1n* region (0.54) than across the rest of the genome (0.01) (Figure 2.1). Genetic distance among accessions for SNPs along chromosome 1 outside the *Inv1n* region shows little evidence of haplotype structure

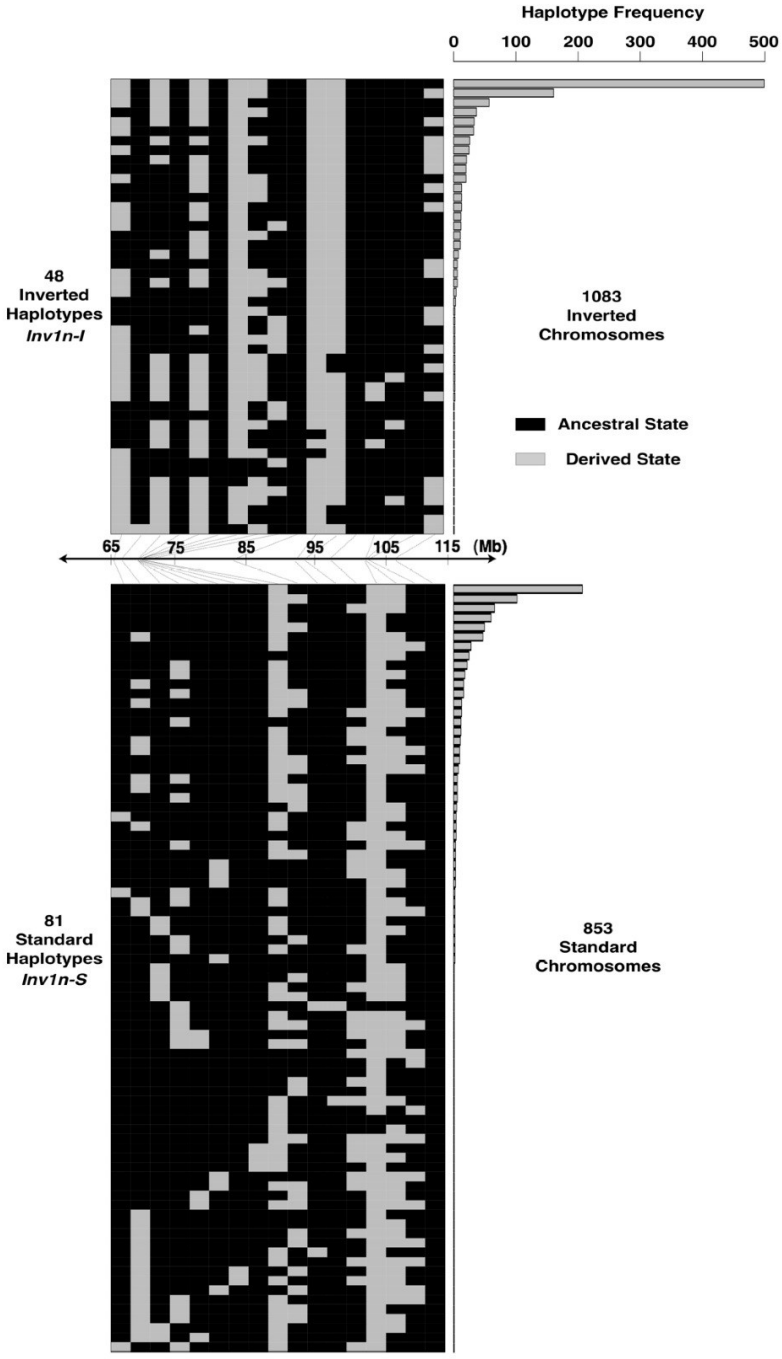


Figure 2.2 Diagram of haplotype diversity in *parviglumis* based on the 17 SNPs within *Inv1n*.

Haplotypes are divided into the two clusters identified by STRUCTURE. Each SNP is represented by either the ancestral state (black) or derived state (gray). The frequency of each of the haplotypes from the inverted (upper panel) and standard (lower panel) arrangements are shown on the right. The middle bar shows the physical position of each of the 17 SNPs inside *Inv1n.*

(Figure 2.3A), while genetic distance for SNPs inside *Inv1n* divides *parviglumis* into two clear haplotypic groups representing *Inv1n-I* and *Inv1n-S* (Figure 2.3B). The *Inv1n-S* cluster includes all taxa of *Zea* and *Tripsacum* investigated, and it is parsimonious to assume that the *Inv1n-I* cluster, present only in *parviglumis* and *mexicana*, represents the derived inverted arrangement (Figure 2.3C). Despite strong differentiation, the two arrangements share polymorphic SNPs (Figure 2.2), even in homozygous individuals unaffected by haplotype phasing (data not shown). Among the 968 *parviglumis* samples, 345 (35.6%) are heterozygous at *Inv1n*, while 369 (38.1%) and 254 (26.3%) are homozygous for the *Inv1n-I* and *Inv1n-S* arrangements, respectively. *Inv1n-I* consists of a smaller number of distinct haplotypes and shows a paucity of rare haplotype variants compared to *Inv1n-S* (Figure 2.2).

Resequencing data from seven loci within *Inv1n* mirror these results (Table 2.1). Four loci (PZA00692, PZA00593, PZA03014 and PZA00146) show distinct haplotype clusters consistent with the SNP genotyping data (data not shown), dividing *parviglumis* into two groups representing *Inv1n-I* and *Inv1n-S.* A comparison of the two groups reveals a higher number of fixed differences, fewer shared derived SNPs, and higher average F_{ST} (0.53 versus 0.05) inside the *Inv1n* region than outside. Average Tajima's *D* of the entire sample is higher inside *Inv1n* (0.58 versus -0.29), and the lack of rare haplotypes on the *Inv1n-I* background observed in the SNP data is reflected in the positive Tajima's *D* at sequences from these chromosomes (Table 2.1). All alleles private to *Inv1n-I* are derived based on *Sorghum* outgroup sequence, but 30% of the alleles private to *Inv1n-S* are ancestral.

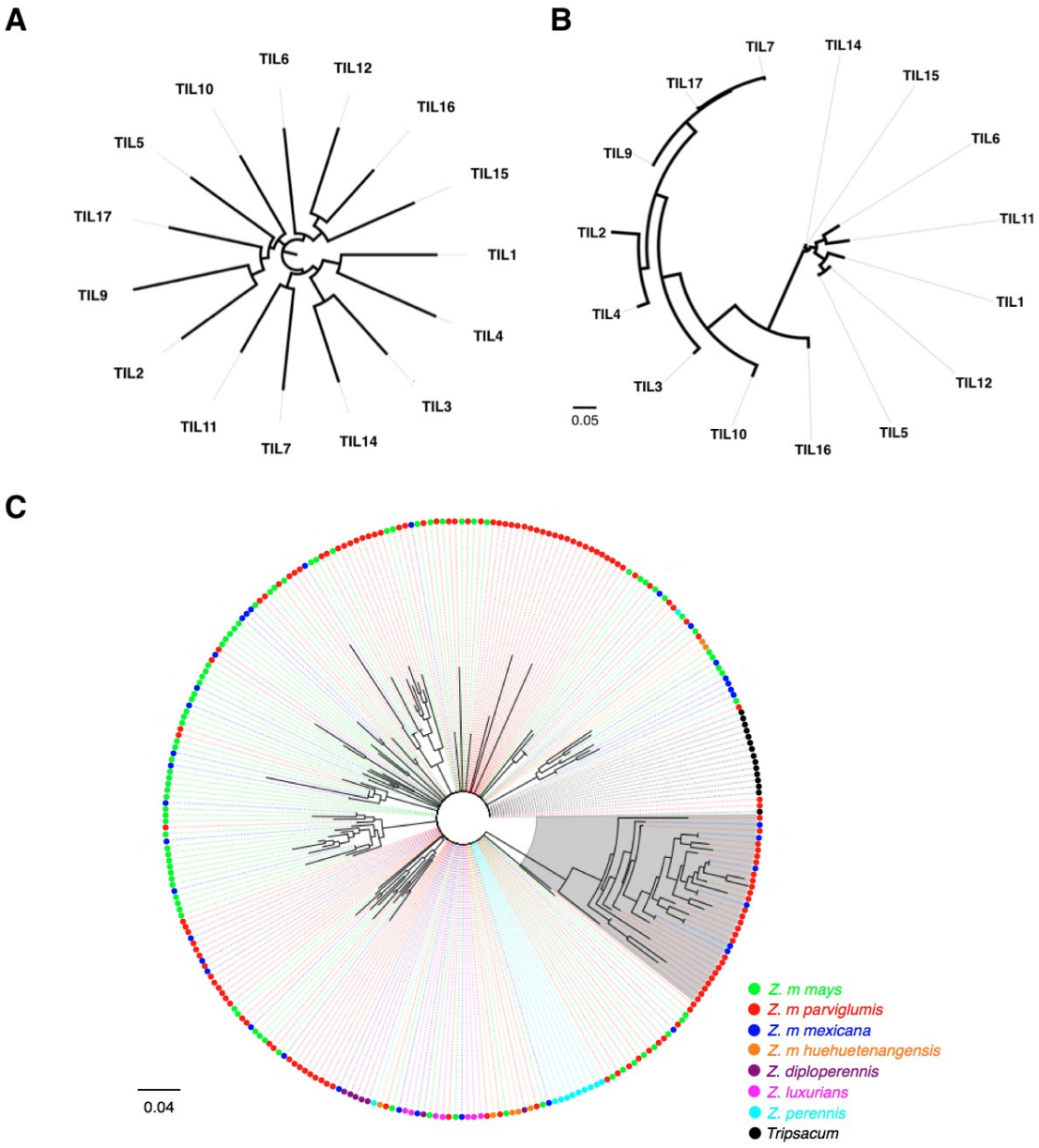


Figure 2.3 Neighbor-joining trees
(A) Neighbor-joining tree for all SNPs outside *Inv1n* using 15 *parviglumis* inbred lines. (B) Neighbor-joining tree for all SNPs inside *Inv1n* using 15 *parviglumis* inbred lines. (C) Neighbor-joining tree for all unique haplotypes in each taxon using all SNPs inside *Inv1n*. The haplotypes in the grey region represent the *Inv1n-I* arrangement.

Table 2.1 Mean (and standard deviation) of summary statistics for seven resequencing loci inside and 88 loci outside *Inv1n*.
The number of loci with an outgroup is listed in parentheses in the # of loci column. The numbers in parentheses in other columns are standard deviations; n: number of samples; L: length of the locus; S_{Sh}: number of shared SNPs between *Inv1n-I* and *Inv1n-S*; S_f: the number of fixed SNPs; S_P: the number of private SNPs; h: number of haplotypes; H: haplotype diversity; θ_π: pairwise difference per base pair.

	# loci	n	L	S_{sh}	S_f	S_p	h	H	θπ	Taj*D*	Fay & Wu's H
Inside (*Inv1n-I*)	7 (6)	14.6 (1.5)	307 (88)	0.3 (0.8)	4.3 (3.1)	2.9 (3.2)	2.3 (1.4)	0.49 (0.41)	0.004 (0.004)	0.37 (1.21)	-0.001 (0.003)
Inside (*Inv1n-S*)						2.9 (2.5)	3.3 (1.7)	0.59 (0.36)	0.004 (0.003)	-0.70 (0.56)	0 (0.003)
Outside (*Inv1n-I*)	88 (68)	13.5 (1.9)	414 (107)	4.7 (4.6)	0.1 (0.9)	4.1 (4.1)	3.9 (1.2)	0.89 (0.21)	0.011 (0.009)	-0.22 (0.65)	-0.003 (0.017)
Outside (*Inv1n-S*)						4.7 (3.6)	4.7 (1.8)	0.88 (0.21)	0.010 (0.007)	-0.34 (0.62)	-0.008 (0.029)

We used multiple approaches to estimate the age of *Inv1n-I* from the resequencing data. Using the MCMC approach of Becquet and Przeworski (2007), which estimates divergence time from patterns of shared polymorphism under an isolation model, divergence was estimated to be ~296,000 generations, with a 95% confidence interval (CI) between 221,000 and 398,000 generations. Assuming a constant rate of substitution of 3×10^{-8} per generation (Clark *et al*., 2005b), we can also calculate divergence time from net differences in nucleotide diversity (Nei and Li, 1979) between *Inv1n-I* and *Inv1n-S*, which gives an estimate of 260,000 generations. Finally, estimates of the time to the most recent common ancestor (TMRCA) (Thomson *et al*., 2000; Hudson, 2007) of the complete sample inside *Inv1n* (308,100 generations, 95% CI of 272,800 ~ 345,600 generations) and of *Inv1n-I* alone (133,200 generations, 95% CI of 96,500 ~ 175,800

generations) are consistent with other methods.

2.3.3 Neutrality tests

Based on standard tests of neutrality, there is limited evidence of selection on *Inv1n*. HKA tests on *Inv1n* did not detect evidence of balancing selection caused by environmental heterogeneity (p-value = 0.46, divergence to diversity ratio = 2.58). MK (p-value = 0.65) and HKA (p-value = 0.15) tests on resequencing data from the *Inv1n-I* arrangement failed to reject a neutral model, and Fay and Wu's H (Fay and Wu, 2000) did not differ markedly between the *Inv1n-I* arrangements or compared to loci outside of *Inv1n* (Table 2.1).

2.3.4 Population frequencies and association analyses

All 33 *parviglumis* populations sampled were polymorphic for both arrangements at the *Inv1n* locus, with a mean *Inv1n-I* frequency of 55%. The frequency of *Inv1n-I* is negatively correlated with altitude ($r^2 = 0.34$) (Figure 2.4; Figure S2.3; Table S2.1), ranging from 90% in the Quenchendio population at an altitude of 653 m to 18.4% in Ahuacatitlan at 1528 m. Consistent with this, *Inv1n-I* occurs at a frequency of only 9.7% in subspecies *mexicana*, which is found at higher altitudes than subspecies *parviglumis* (mean altitude of 2091 m versus 1087 m for our *mexicana* and *parviglumis* samples). The most common *Inv1n-I* haplotype makes up 46% (499/1083) of *parviglumis* chromosomes with the *Inv1n-I* variant, and does not vary significantly in frequency among populations

(χ^2 = 2.27, df = 32, p-value = 1).

Using a model-based approach (Coop *et al*., 2010) to control for population structure, we examined the association between *Inv1n-I* frequency and 22 environmental variables (Table S2.3). Among environmental variables, altitude was most strongly correlated with *Inv1n-I* frequency and consistently obtained the highest Bayes factors among runs (mean 136). High Bayes factors were also observed for other bioclimatic variables, including temperature (mean temperature of driest quarter; mean Bayes factor

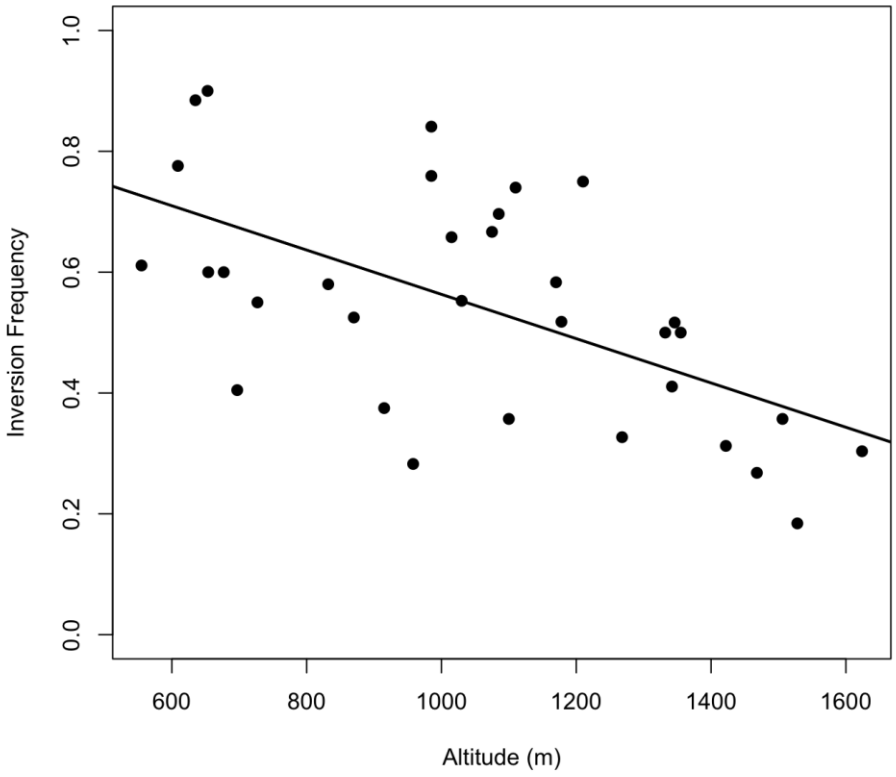


Figure 2.4 *Inv1n-I* frequency in *parviglumis* populations is negatively correlated with altitude.
Each point corresponds to a population, with its altitude on the x-axis and *Inv1n-I* frequency on the y-axis.

49) and precipitation (precipitation of driest month; mean Bayes factor 48). Genome-wide analysis of all SNPs produced mean Bayes factors for association with altitude ranging from 0.29 to 1048 (Table S2.3). SNPs in the inversion are tenfold enriched in the top 5% tail of Bayes factors for altitude, and more than twentyfold enriched in the top 1% tail, strongly suggesting a link between *Inv1n* and altitude (Figure 2.5).

We also used a mixed linear model analysis to test for associations between *Inv1n-I* and phenotypic data from the same populations (Weber *et al*., 2008). *Inv1n-I* appears to be associated with the percentage of male internodes (PSIN) (p-value = 0.0055, r^2 = 0.024), percentage of staminate spikelets (STAM) (p-value = 0.0069, r^2 = 0.023), culm diameter (CULM) (p-value = 0.0137, r^2 = 0.011) and leaf number (LFNM) (p-value = 0.0232, r^2 = 0.010) (Table S2.4), but none of the associations are significant after Bonferroni correction for phenotypes tested and effect sizes for all phenotypes are very small. In addition to testing the inversion as a single locus, we also investigated associations between individual SNPs and phenotypes. For both PSIN and STAM, none of the 17 SNPs in *Inv1n* were among the 1% of SNPs most strongly associated with the two phenotypes. However, SNPs in *Inv1n* were enriched in the 1% tail of p-values for both CULM (15x enrichment) and LFNM (3x). None of the SNPs were significantly associated with PSIN or LFNM at a false discovery rate (FDR) of 5%, while four SNPs outside of the inversion were significantly associated with STAM. Of the seven SNPs significantly associated with CULM at an FDR of 5%, four (PZA00263.14, PZD00077.7, PZA00692.5, PZA03014.24) are inside *Inv1n*.

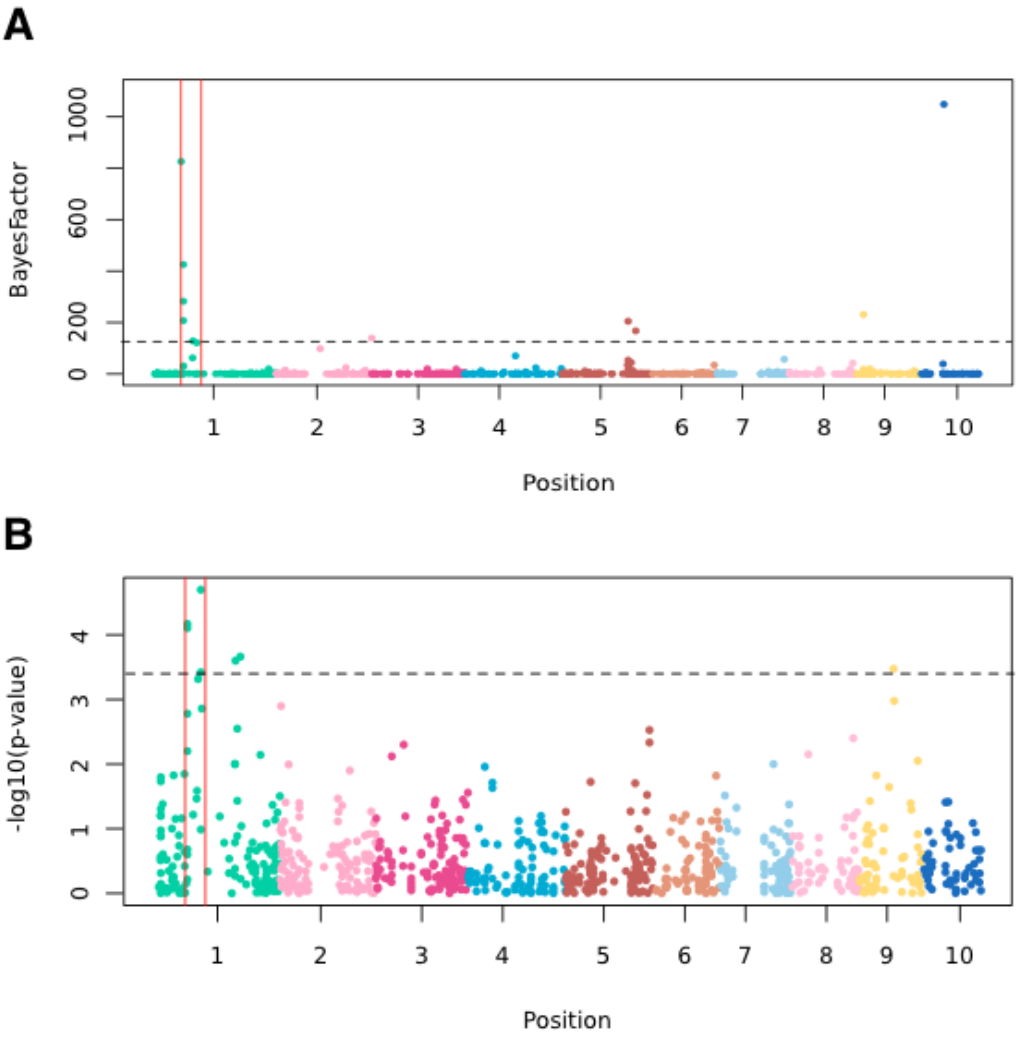


Figure 2.5 (A) Bayes factors for correlation between allele frequencies and altitude in 33 natural *parviglumis* populations. (B) Association between all SNPs and culm diameter.

In (A), *Inv1n* is indicated by red vertical lines. The 99th percentile of Bayes factors distribution is indicated by horizontal lines. Chromosomes 1 to 10 are plotted in order and in different colors. In (B), SNPs significant at 5% FDR are above the dashed line.

2.4 Discussion

Using a genome-wide set of SNPs in a large panel of wild and domesticated *Zea*, we provide evidence of an ~50-Mb inversion on the short arm of chromosome 1 of subspecies *parviglumis* and *mexicana*. While our cytological data are not directly diagnostic for an inversion, population genetic data preclude alternative explanations. For example, the dramatic reduction in haplotype number in the *Inv1n* region (Figure 2.1) could be indicative of a selective sweep (Kim and Nielsen, 2004; Nielsen *et al*., 2005; McVean, 2007). However, the largest sweep identified in maize to date is only 1.1 Mb (Tian *et al*., 2009), and both the age of the inversion and common tests for departures from neutrality do not provide evidence of strong selection. Another alternative explanation would be the presence of strong negative interactions between distantly linked loci, potentially due to synthetic lethality (Boone *et al*., 2007). Such interactions should not generate extended patterns of elevated LD among intervening SNPs, as crossing-over among haplotypes not carrying alleles involved in the negative interaction should not be affected. Both selective sweeps and negative interactions are inconsistent with the presence of only two major haplotypes in the *Inv1n* region and fail to explain the clinal variation in haplotype frequencies seen at *Inv1n-I*.

To our knowledge, the only prior evidence for *Inv1n* is a report of high LD and high F_{ST} from a much smaller sample of *parviglumis* (Hufford *et al*., 2012), but a number of other large inversions have been previously reported in *mays* and its wild relatives (Ting, 1965; Maguire, 1966; Ting, 1967; Kato Y., 1975; Ting, 1976). These include an ~50-Mb

inversion on the long arm of chromosome 3 in *Zea luxurians* (Ting, 1965) and an ~35-Mb inversion that covers most of the short arm of chromosome 8 in both *mays* (McClintock, 1960) and *mexicana* (Ting, 1976). While some of these inversions were experimentally induced (McClintock, 1931; Morgan, 1950), several have also been identified in natural populations of multiple taxa (Kato Y., 1975; Ting, 1976).

One of the factors that may limit the geographic spread of large inversions is the potential fitness cost of crossing over. The frequency of chromosome loss is dependent on the inversion size and efficiency of synapsis over the inverted region (Burnham, 1962; Maguire and Riess, 1994; Lamb *et al*., 2007). When gene density is low, such as in pericentromeric regions, or there is a lack of continuous homology, chromosomes will often synapse in a non-homologous manner without recombination (McClintock, 1933). In maize, for example, an inversion on the long arm of chromosome 1 similar in size to *Inv1n* (19 cM) was seen to undergo homologous pairing in only ~1/3 of cases (Maguire, 1966). Since *Inv1n* is located in a pericentromeric region with low gene density and covers a short genetic distance (roughly 2-13 cM), we anticipated that it would rarely pair and recombine with a non-inverted chromosome. Our data are consistent with these arguments. We observed repressed recombination around *Inv1n*, and no cytological evidence of crossing-over in inversion heterozygotes. SNP data indicate no deviations from expected Hardy-Weinberg genotype frequencies at *Inv1n*, and we see no obvious evidence of effects on fertility. Given these observations, we suspect that inversion polymorphisms may be relatively common in natural plant populations, especially in regions of the genome with low recombination rates such as pericentromeres. Low

recombination has also been offered as an explanation for the lack of underdominance in many pericentromeric inversions in *Drosophila* (Coyne *et al*., 1993). As dense genotyping becomes more cost effective, we predict that numerous common inversions will be identified in natural populations of *Zea* and other organisms.

2.4.1 Origin and age of *Inv1n*

Our evidence suggests that *Inv1n-I* is the derived, inverted arrangement. *Inv1n-I* is not found in *Tripsacum* or *Zea* taxa except for *parviglumis* and *mexicana* (Figure 2.3C), and, unlike *Inv1n-S*, all SNPs private to *Inv1n-I* are derived in resequencing data. Both SNP and resequencing data show strong differentiation between the two arrangements (Figure 2.1; Table 2.1). Multiple methods of estimating the age of the *Inv1n-I* haplotype point to an origin ~300,000 generations ago. This predates both the split between *mexicana* and *parviglumis* and the split between *Zea luxurians* and *Zea mays*, and is similar to the estimated age of divergence of most species in the genus *Zea* (Ross-Ibarra *et al*., 2009). Several considerations suggest that these numbers are plausible. First, the proportion of SNPs shared between *parviglumis* and *mexicana* on chromosome 1 does not differ inside or outside of *Inv1n* (15/17 versus 134/139, Fisher's exact test p-value = 0.48), suggesting that the presence of the inversion in both subspecies is likely due to shared ancestral polymorphism rather than recent gene flow. Second, while the estimated age of *Inv1n-I* is similar to the estimated divergence of species, other species in *Zea* have narrow distributions (Fukunaga *et al*., 2005) and presumably small effective population sizes, increasing the potential for loss of variants at low-frequency in the ancestral

population. Third, *Inv1n-I* could consist of multiple independent inversions, similar to the inversion polymorphisms identified in the white-throated sparrow (Thomas *et al*., 2008). In this case, estimates of the age of the inversion would be biased upwards, as each inversion would have arisen independently on distinct backgrounds. Such a scenario might also explain the observation of shared polymorphisms between *Inv1n-I* and *Inv1n-S*. Our data cannot distinguish the number of independent inversions in the region, however, which would instead require analysis of progeny derived from crosses of multiple individuals homozygous for different haplotypes of *Inv1n-I* and a more dense set of markers.

While small effective population size may explain the absence of *Inv1n-I* from other taxa in *Zea*, its complete absence in our sample of 1573 *mays* requires additional explanation. Sampling alone is unlikely to play a role, as the vast majority of our *mays* accessions are landraces, collected from across the Americas, including accessions collected within the range of *parviglumis* and *mexicana*. Estimates of the domestication bottleneck and observed levels of diversity in domesticated *mays* (Tenaillon *et al*., 2004; Wright *et al*., 2005) also suggest that drift during domestication is not a compelling explanation, especially given that *Inv1n-I* occurs at frequencies of up to 90% in the lowland areas where domestication is thought to have occurred (Matsuoka *et al*., 2002; Piperno *et al*., 2009; van Heerwaarden *et al*., 2011). We speculate instead that *Inv1n-I* may have been selected against in domesticated *mays*. Our association analysis provides limited evidence in support of this idea, as *Inv1n-I* is negatively associated with culm width in *parviglumis,* while domesticated *mays* has more robust culms than its wild

progenitor (Briggs *et al*., 2007).

2.4.2 Selection on *Inv1n*

While standard tests of neutrality do not provide evidence for selection, there is reason to believe that *Inv1n* is not evolving neutrally. First, the *Inv1n-I* arrangement is widely distributed, segregating in all 33 populations investigated; only two SNPs on chromosome 1 are also polymorphic in all populations. Second, pairwise F_{ST} inside *Inv1n* appears uncorrelated to pairwise F_{ST} genome-wide ($r^2 = 0.04$; Figure S2.4), suggesting that the frequency of *Inv1n* is not entirely due to isolation by distance. Third, even after correcting for population structure, *Inv1n-I* frequency is associated with a number of environmental variables (Table S2.3), including a strong altitudinal cline (Figures 2.4; Figure S2.3). Latitudinal (Anderson *et al*., 2005; Santos *et al*., 2005; Umina *et al*., 2005) and altitudinal clines (Levitan, 2001) are commonly observed for inversion polymorphisms, and are often thought to be related to temperature adaptation (Levitan, 2001; Umina et al., 2005; Balanyá *et al*., 2006). Fourth, the *Inv1n-I* arrangement is, to our knowledge, the first inversion in *Zea* shown to be associated with phenotypic differences (Table S2.4). These include culm diameter, a trait which differentiates maize from teosinte (Briggs *et al.,* 2007), and tassel morphology (Table S2.4), which is known to differ between *parviglumis* and *mexicana* (Doebley, 1983) and between lowland and highland maize (Anderson, 1946; Bretting and Goodman, 1989). Fifth, the lack of rare variants and the high frequency of the most common *Inv1n-I* haplotype (Figure 2.2) suggest that this haplotype may have recently risen to high frequency due to a partial

sweep. The observed lack of rare variants is especially striking given the genome-wide pattern of an excess of low frequency variants (Table 2.1), an observation reported in multiple studies (Tenaillon *et al*., 2004; Wright *et al*., 2005; Moeller *et al*., 2007; Ross-Ibarra *et al*., 2009). While the most common haplotype at *Inv1n-I* does not show signs of extended homozygosity beyond the borders of the inversion (Figure S2.5) as might be expected if it has been recently swept to higher frequency, the nearest flanking SNPs are 1.1 and 14.6 Mb distant and our power to detect an extended haplotype is low. Sixth, the absence of *Inv1n-I* from domesticated maize, in spite of recurrent gene flow from both *parviglumis* and *mexicana* (Wilkes, 1967; Fukunaga *et al*., 2005; Ross-Ibarra *et al*., 2009; van Heerwaarden *et al*., 2011), suggests that the inverted arrangement was selected against at some point during *mays* domestication or breeding. Finally, we note that, aside from strong divergence between chromosomes of different arrangements, selection may be difficult to detect in diversity data from inversions of an age similar to ours (Guerrero *et al*., 2012).

Selection may act on inversions because of the fitness consequences of the structural rearrangement itself or of adaptive alleles at loci inside the inversion (Kirkpatrick and Barton, 2006). While these models predict somewhat different patterns of diversity (Guerrero *et al*., 2012), our SNP genotyping data is of insufficient density to distinguish between them. Regardless of the model of selection, the observed altitudinal cline and absence in domesticated *mays* suggest that *Inv1n-I* is not ubiquitously adaptive. Even in low altitude populations where it seems to be favored, a large inversion such as *Inv1n* may have captured several recessive deleterious alleles, effectively preventing its fixation

(Kirkpatrick and Barton, 2006).

2.5 Supporting Information

2.5.1 Supplementary Figures

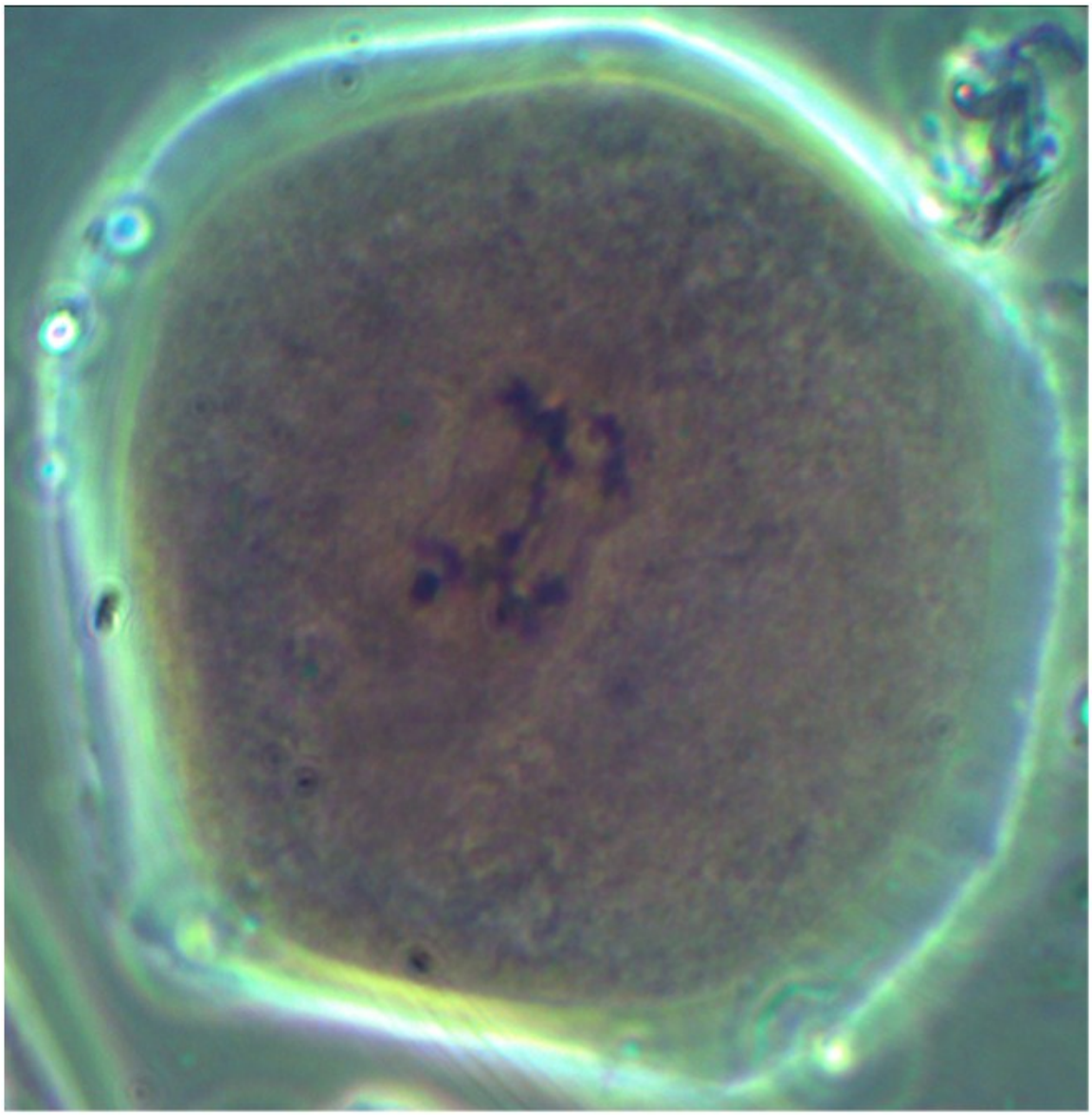


Figure S2.1 An anaphase I bridge in a plant heterozygous for *Inv1n*.
Such bridges were rare, observed in only ~4% of the meiocytes undergoing anaphase I.

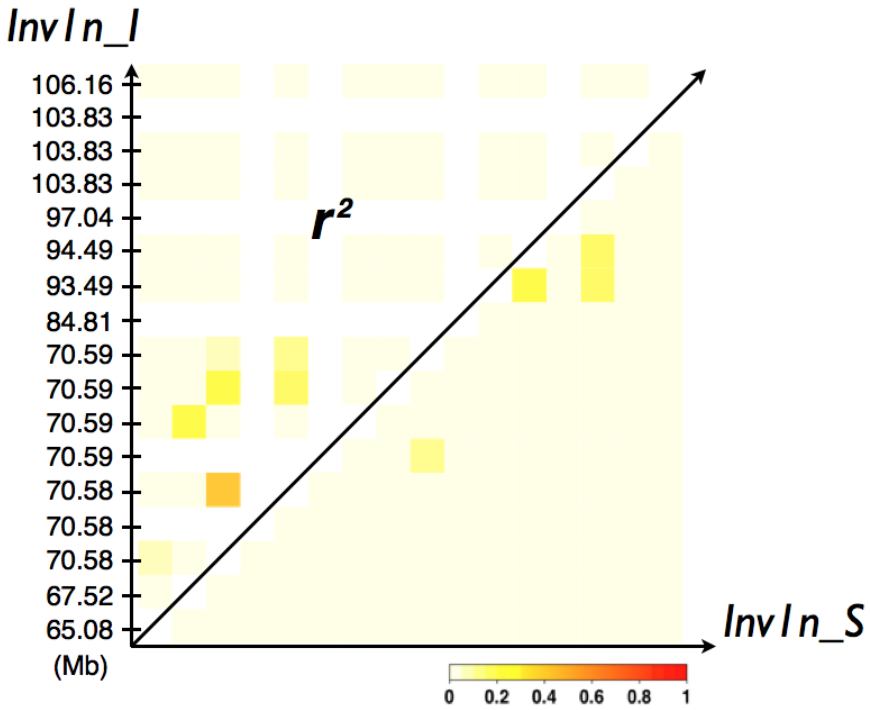


Figure S2.2 LD (r^2) among the 17 SNPs inside *Inv1n* in *parviglumis.*
The physical positions of the 17 SNPs are shown on the left. The upper triangle represents LD in *Inv1n-I*, and *Inv1n-S* is shown in the lower triangle.

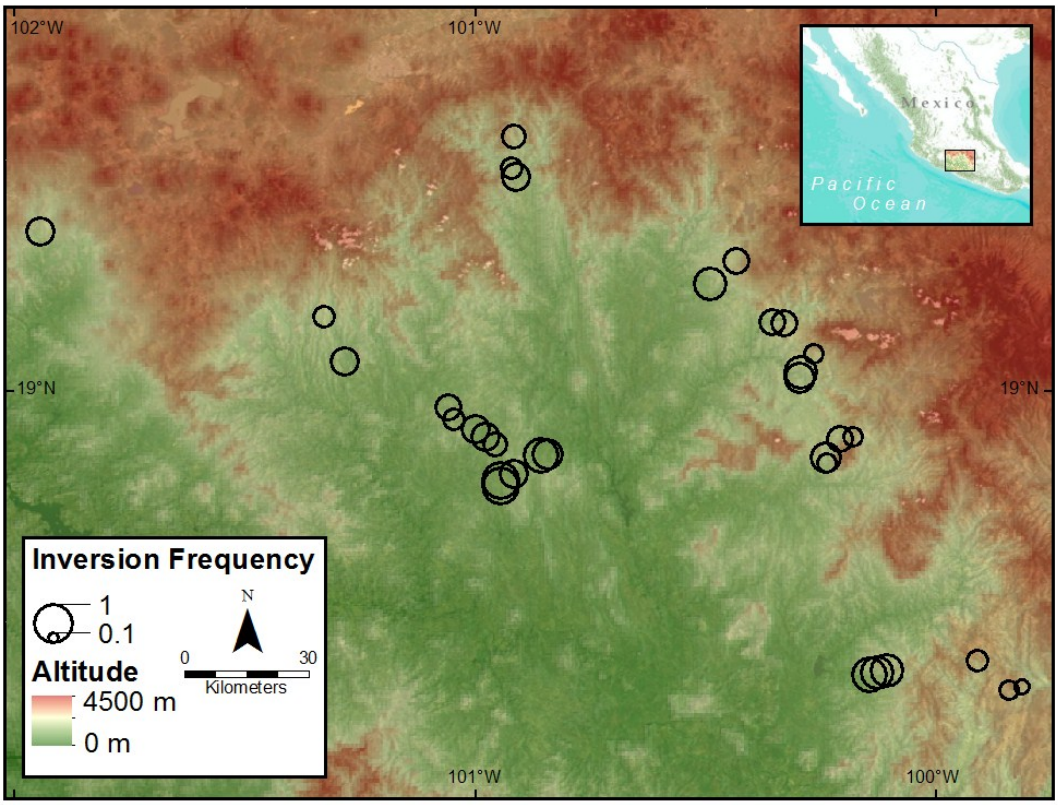


Figure S2.3 Geographic distribution of the 33 *parviglumis* populations.
The size of the circle is proportional to the *Inv1n* frequency, and color represents elevation. The study area in Mexico is shown in the inset.

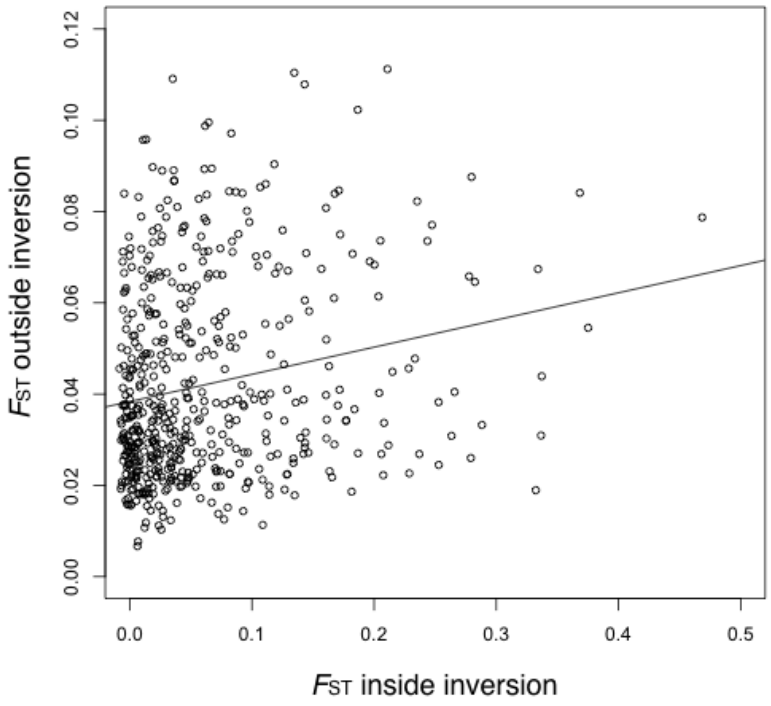


Figure S2.4 Pairwise F_{ST} among 33 *parviglumis* natural populations at SNPs inside *Inv1n* compared to SNPs outside *Inv1n.*

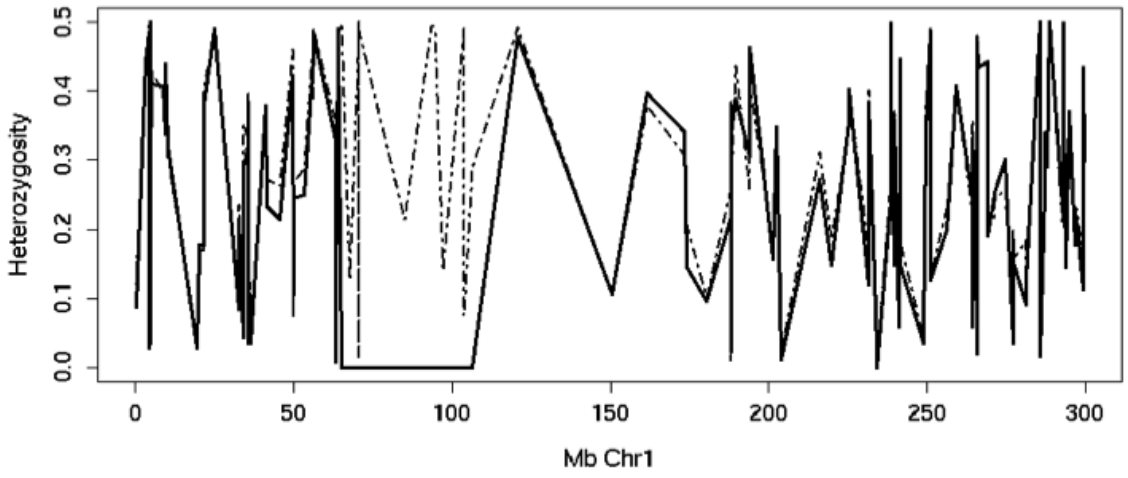


Figure S2.5 Expected SNP heterozygosity across chromosome 1 for all *parviglumis* (dashed line) and the most common *Inv1n-I* haplotype (solid line).

2.5.2 Supplementary Tables

Table S2.1 Location of the 33 *parviglumis* study populations with mean per-SNP values of summary statistics.
N: sample size; F: frequency of *Inv1n-I*.

Population	State/ Province	Latitude	Longitude	Altit-ude	N	F	HW E	θ_H	π	Taj*D*	Fay/ Wu H
Crustel	Guererro	18.383	-100.145	985	22	0.84	0.48	0.32	0.26	1.31	-0.06
Amates 1	Guererro	18.388	-100.128	1110	25	0.74	0.47	0.31	0.28	1.36	-0.03
Amates 2	Guererro	18.394	-100.108	1210	24	0.75	0.39	0.30	0.28	1.31	-0.02
Iguala	Guererro	18.414	-99.909	1506	28	0.36	0.96	0.31	0.28	1.51	-0.03
Rincon	Guererro	18.350	-99.841	1624	28	0.30	0.20	0.31	0.27	1.68	-0.04
Ahuacatitlan	Guererro	18.356	-99.814	1528	19	0.18	0.59	0.29	0.27	1.43	-0.02
Huetamo 1	Michoacan	19.063	-101.283	832	25	0.58	0.63	0.31	0.25	1.28	-0.06
Puerto 1	Michoacan	18.963	-101.058	870	20	0.53	0.66	0.32	0.26	1.26	-0.06
Zapote	Michoacan	18.938	-101.048	915	24	0.38	0.59	0.31	0.26	1.33	-0.05
Puerto 2	Michoacan	18.916	-101.000	727	20	0.56	0.58	0.31	0.24	1.12	-0.07
Huetamo 2	Michoacan	18.900	-100.979	677	20	0.60	0.46	0.30	0.25	1.09	-0.05
Cuirindalillo	Michoacan	18.883	-100.957	697	21	0.40	0.69	0.31	0.25	1.03	-0.06
Crucero	Michoacan	18.794	-100.946	653	25	0.90	0.02	0.30	0.23	1.17	-0.07
Quenchendio	Michoacan	18.805	-100.946	635	26	0.88	0.51	0.31	0.25	1.09	-0.06
Potrero	Michoacan	18.820	-100.916	654	20	0.60	0.85	0.30	0.26	1.12	-0.04
Crucita	Michoacan	18.858	-100.857	609	29	0.78	0.56	0.31	0.25	1.25	-0.06
Guayabo	Michoacan	18.862	-100.844	555	27	0.61	0.12	0.31	0.25	1.23	-0.06
Toluca 1	Mexico	18.899	-100.181	1422	24	0.31	0.74	0.31	0.27	1.33	-0.04
Toluca 2	Mexico	18.895	-100.209	1355	23	0.50	0.86	0.31	0.28	1.22	-0.03
Toluca 3	Mexico	18.854	-100.239	1015	19	0.66	0.82	0.32	0.27	1.06	-0.05
Salitre-Monte	Mexico	18.842	-100.238	958	23	0.28	0.23	0.31	0.27	1.17	-0.04
Taretan	Michoacan	19.344	-101.944	1170	18	0.58	0.90	0.29	0.21	0.98	-0.08
Los Guajes	Michoacan	19.231	-100.491	985	27	0.76	0.65	0.31	0.27	1.36	-0.04
Norte	Michoacan	19.281	-100.434	1332	27	0.50	0.85	0.30	0.27	1.33	-0.03
Zuluapan 1	Mexico	19.148	-100.355	1178	28	0.53	0.46	0.31	0.28	1.33	-0.03
Zuluapan 2	Mexico	19.146	-100.329	1346	30	0.52	0.73	0.31	0.27	1.30	-0.04
Zacazonapan 1	Mexico	19.079	-100.266	1468	28	0.27	0.99	0.30	0.26	1.33	-0.04

Zacazonapan 2	Mexico	19.039	-100.295	1085	28	0.70	0.71	0.32	0.27	1.38	-0.05
El Puente	Mexico	19.029	-100.296	1075	24	0.67	0.76	0.31	0.27	1.30	-0.04
Queretanillo	Michoacan	19.551	-100.918	1342	28	0.41	0.32	0.32	0.26	1.33	-0.06
Temascal 1	Michoacan	19.483	-100.921	1100	28	0.36	0.80	0.32	0.27	1.33	-0.05
Temascal 2	Michoacan	19.464	-100.912	1030	19	0.55	0.27	0.31	0.27	1.18	-0.04
Casa Blanca	Michoacan	19.161	-101.329	1268	26	0.33	0.00	0.33	0.26	1.40	-0.07

Table S2.2 Counts of anaphase and telophase pollen meiocytes showing dicentric bridges or normal segregation during meiosis

	Line	Normal	Bridge	**Sum**
1	B73 x TIL5	17	0	**17**
2	B73 x TIL5	45	3	**48**
3	OH43 x TIL11	48	2	**50**
4	OH43 x TIL11	36	1	**37**
5	OH43 x TIL11	4	0	**4**
6	OH43 x TIL11	17	1	**18**

Table S2.3 Mean Bayes factors for all environmental variables and inversion as single marker, all the SNPs in *Inv1n* and all SNPs.
T: temperature.

	Inversion	SNPs in *Inv1n*	All SNPs
Longitude	0.36	0.69	1020.37
Latitude	1.64	0.87	34.59
Altitude	136.37	124.63	5.86
Annual Mean T	18.93	12.57	5.75
Mean Diurnal T Range	0.82	4.35	28.61
Isothermality	1.26	0.84	2.77
T Seasonality	0.92	0.87	2.15
Max T of Warmest Month	22.46	16.42	5.76
Min T of Coldest Month	11.47	5.69	7.43
T Annual Range	1.90	9.75	28.75
Mean T of Wettest Quarter	21.04	17.20	9.33
Mean T of Driest Quarter	48.98	26.87	3.35
Mean T of Warmest Quarter	21.27	13.83	5.30
Mean T of Coldest Quarter	17.19	10.34	6.66
Annual Precipitation	0.88	1.26	2.69
Precipitation of Wettest Month	0.33	0.53	19.25
Precipitation of Driest Month	47.29	17.46	2.40
Precipitation Seasonality	10.54	3.67	2.24
Precipitation of Wettest Quarter	0.46	0.70	2.89
Precipitation of Driest Quarter	34.48	17.86	4.53
Precipitation of Warmest Quarter	0.44	0.51	1.73
Precipitation of Coldest Quarter	5.21	2.45	1.43

Table S2.4 Results of association analysis.
P-value, marker r^2, *a* (genotypic value of inversion homozygote) and *d* (genotypic value of heterozygote) are based on analysis of *Inv1n* as single marker. Number of significant SNPs at FDR 5% is reported separately for all SNPs and for SNPs inside *Inv1n*.

	Inversion as single marker				Association analysis for all SNPs	
Trait	p	marker r^2	a	d	Significant SNPs at FDR 5%	Significant SNPs in inversion at FDR 5%
Blade Length[a]	0.7065	0.001	-0.33	0.15	0	0
Culm diameter	0.0137	0.011	-0.55	0.54	7	4
Days to Pollen	0.2595	0.004	-0.50	0.69	0	0
Days to Silk	0.4831	0.002	-0.33	0.54	0	0
Female ear length[b]	0.3304	0.005	1.18	-0.80	0	0
Fruitcase compression[b]	0.0662	0.007	-0.11	0.03	0	0
Fruitcase length[b]	0.1393	0.008	0.12	-0.03	0	0
Fruitcase weight	0.5961	0.001	0.00	0.00	0	0
Lateral branch internode number[a]	0.663	0.001	-1.05	0.56	0	0
Lateral Branch Length[a]	0.8508	0.001	-0.02	-0.06	0	0
Lateral inflorescence branch number[a]	0.5653	0.002	-0.62	0.42	0	0
Lateral inflorescence length	0.3641	0.005	-0.65	0.94	0	0
Leaf Number	0.0232	0.01	-0.12	0.68	0	0
Leaf Width	0.1243	0.005	-0.13	0.11	2	0
Maize Introgressed	0.5062	0.002	-0.01	0.00	4	0
Mean lateral branch internode length[a]	0.7455	0.001	-0.27	0.32	0	0
Number of Barren nodes	0.7907	0.001	0.00	0.01	0	0
Number of female cupules[b]	0.7164	0.001	0.03	-0.13	0	0
Number of female internodes[b]	0.8604	0.001	0.06	-0.09	0	0
Oil Content, Wet	0.7653	0.001	0.01	0.04	0	0
Paired Spikelets	0.5523	0.002	0.00	0.00	5	0

Pedicellate Spikelet	0.1717	0.006	0.00	0.01	15	1
Percent of Male Internodes in the lateral inflorescence	0.0069	0.023	-2.62	2.75	0	0
Percent staminate spikelets in the lateral inflorescence	0.0055	0.024	-1.96	2.08	4	0
Plant Height	0.1175	0.005	-1.21	6.42	0	0
Polystichous	0.1162	0.007	-0.01	0.02	2	0
Prolificacy[a]	0.3762	0.002	-0.41	0.28	0	0
Proportion of female cupules[b]	0.2483	0.007	0.02	-0.01	0	0
Proportion of female ear length[b]	0.3717	0.005	0.02	-0.01	0	0
Proportion of female internodes[b]	0.2604	0.007	0.02	-0.01	0	0
Protein Content	0.5198	0.002	-0.22	0.07	0	0
Sexual Identity of the Lateral Inflorescence	0.294	0.003	-0.02	0.05	0	0
Sheath Length	0.9239	0	-0.03	0.04	0	0
Starch Content[c]	0.4999	0.002	0.21	-0.12	0	0
Tassel branch number	0.1425	0.01	-5.80	0.64	0	0
Tiller number	0.2321	0.004	0.55	0.58	0	0
Yoked Cupules	0.7034	0.001	0.00	0.00	21	0

[a] On the second lateral branch from the top of the plant
[b] In the basal ear
[c] Adjusted to percent dry matter using a pooled moisture content estimate

CHAPTER 3

COMPARATIVE ANALYSES IDENTIFY THE CONTRIBUTIONS OF EXOTIC DONORS TO DISEASE RESISTANCE IN A BARLEY EXPERIMENTAL POPULATION

Introgression of novel genetic variation into breeding populations is frequently required to facilitate response to new abiotic or biotic pressure. This is particularly true for the introduction of host pathogen resistance in plant breeding. However, the number and genomic location of loci contributed by donor parents are often unknown, complicating efforts to recover desired agronomic phenotypes. We examine allele frequency differentiation in an experimental barley breeding population subject to introgression and subsequent selection for *Fusarium* head blight resistance. Allele frequency differentiation between the experimental population and the base population identifies three primary genomic regions putatively subject to selection for resistance. All three genomic regions have been previously identified by quantitative trait locus (QTL) and association mapping. Based on the degree of identity by state relative to donor parents, putative donors of resistance alleles are also identified. The successful application of comparative population genetic approaches in this barley breeding experiment suggests that the approach could be applied to other breeding populations that have undergone defined breeding and selection histories, with the potential to provide valuable information for genetic improvement.

3.1 Introduction

Experimental evolution studies have long been valued as a means of gaining insight into the rate and degree of response to selection (cf. Burke and Rose, 2009). Recent studies have demonstrated that when combined with single-nucleotide polymorphism (SNP) genotyping or resequencing and comparative population genetic approaches, experimental populations can provide a rapid means of identification of loci responding to selection (Burke *et al.*, 2010; Turner *et al.*, 2011; Orozco-Terwengel *et al.*, 2012). The approaches have been successfully applied to a number of systems, from experimental populations of microbes exposed to high temperature regimes (Tenaillon *et al.*, 2012), to *Drosophila* populations selected for longevity (Teotónio *et al.*, 2009; Burke *et al.*, 2010). These studies demonstrate the feasibility of identifying adaptive mutations in experimental populations.

While plant breeding populations are themselves highly successful long-term experimental evolution studies, there is a long-standing tradition of developing experimental populations to supplement traditional breeding approaches (Harlan and Martini, 1929; Suneson, 1956; Dudley and Lambert, 2004). These experimental populations explore both the potential for local adaptation (Allard *et al.*, 1972) and the genetic basis of response to selection (Clegg *et al.*, 1972; Weir *et al.*, 1972). Experimental plant populations have also been employed to test the role of genetic incompatibility in the formation of hybrid species (Rieseberg *et al.*, 1996) and the potential for trait introgression in cultivated species (Lin *et al.*, 1995; Jiang *et al.*, 2000).

Historically, uses of population genetic approaches to identify genetic changes in experimental populations have been limited to a handful of markers, limiting inference regarding the loci contributing to genetic differentiation to single markers representing large genomic regions (Allard *et al.*, 1972; Rieseberg *et al.*, 1996). Recently the availability of high density SNP genotyping or resequencing data has provided the potential to identify precise genomic regions that have been under selection. Strong directional selection has the potential to produce populations with dramatic differentiation in allele frequency, a potential signal of selection (Lewontin and Krakauer, 1973), which could be detectable either by SNP genotyping (Teotónio *et al.*, 2009) or resequencing (Burke *et al.*, 2010; Turner *et al.*, 2011). Patterns of linkage disequilibrium (LD) (Sabeti *et al.*, 2002) or changes in patterns of identity by state (IBS) (Albrechtsen *et al.*, 2010a) can also suggest recent selection.

In the present study, we report a population genetic examination of an experimental barley breeding population developed in response to epidemic levels of the fungal pathogen *Fusarium graminearum*, the causal agent of *Fusarium* head blight (FHB). The prevalence and spread of this pathogen increased in the Midwestern U.S. in the early 1990s (McMullen *et al.*, 1997) and revealed limited genetic variation for disease resistance among existing barley cultivars and the need to introduce novel variation for resistance into breeding programs. During the 35 years preceding the FHB outbreak, the University of Minnesota barley breeding program made use of relatively closed pedigrees in an advanced cycle breeding scheme (Rasmusson and Phillips, 1997) that primarily focused on crosses among elite lines from within the breeding population (Condón *et al.*,

2009). After the outbreak, numerous exotic sources of elite lines with known FHB resistance and reduced deoxynivalenol (DON) mycotoxin (which is produced by the pathogen) concentration were evaluated and introduced as parents in the breeding population to enhance FHB resistance (Smith *et al.*, 2013).

Novel adaptive mutations, including resistance to a pathogen, are likely to be rare in breeding populations, as mutation frequency is dependent on effective population size. Introduction of resistance alleles from standing variation outside the breeding population is the primary mechanism to increase disease resistance (Fetch *et al.*, 2003). We report the genetic effects of introgression from a diverse set of 13 barley lines carrying FHB resistance into an existing breeding population. The immediate goal of this experimental population (hereafter referred to as the Reopened population) is to provide substantial improvement in resistance to FHB infection. The Reopened population was compared to a contemporaneous sample of breeding lines from the primary Minnesota breeding population, never subject to introgression of FHB resistant parents and maintained under a closed pedigree (hereafter referred to as the Closed population) permitting identification of loci potentially subject to selection for FHB resistance in the Reopened population. We demonstrate that comparative population genetic approaches applied to this experimental population provide a complementary approach to QTL and association mapping methods for identifying loci that underlie important phenotypes.

3.2 Materials and Methods

3.2.1 Plant Materials

Breeding lines for this study were derived from the six-row malting barley breeding program at the University of Minnesota, which began in the early 1900s. The genetic base of this breeding population is quite narrow with ~50% of the six-row germplasm in North America tracing to five ancestors (Martin *et al.*, 1991). In the early 1990s, the original advanced cycle breeding strategy was maintained for part of the breeding program, while a new strategy that introduced exotic sources of FHB resistance was implemented in parallel. This resulted in two parallel breeding populations within the breeding program with different breeding histories. To compare these two populations, we created two panels of 120 breeding lines that were representative of the two populations. The "Closed" panel, is comprised of lines from the advanced cycle breeding program (elite x elite) with a relatively closed pedigree, i.e., few new founders were introduced after 1958 when the strategy was initiated (Condón *et al.*, 2008; Condón *et al.*, 2009). The "Reopened" panel is comprised of lines from families derived from the introduction of 13 new donors to the Closed population in response to the FHB epidemic (Figure 3.1). The lines in each panel were selected from a period of transition between advanced cycle breeding and introduction of disease resistant parents (2003 - 2007), such that we could adequately sample both breeding populations. Lines in each panel were selected to maximize the number of families represented within each population; the Closed panel sampled 32 (94%) of 34 families in the Closed population that advanced to

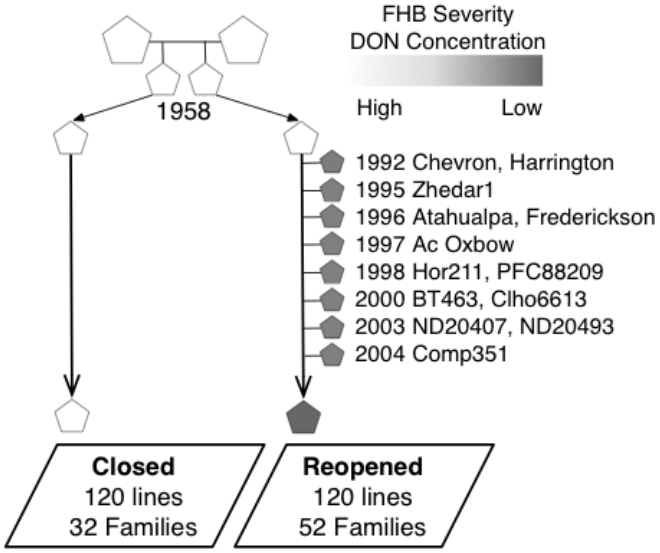


Figure 3.1 Breeding history of the Closed and Reopened populations.
Filled shapes designate individuals carrying FHB resistance and/or reduced DON accumulation. The Reopened population acquired reduced FHB severity and DON concentration through the introgression from 13 donor lines between 1992 and 2004.

preliminary yield trials, and the Reopened panel sampled 52 (87%) of 60 families in the Reopened population. Lines selected for both panels were based on seed or DNA availability, with preference given to lines with malting quality data.

The typical development of breeding lines begins with a cross between two parents in the fall followed by self-pollination of the F_1 in a greenhouse in the winter. Parents were typically selected after they have been evaluated for two years in yield trials based on agronomic performance and acceptable malting quality for the Closed population and additionally for FHB resistance in the Reopened population. F_2 plants were grown in the

field and advanced by single seed descent to the F_4 generation without selection. In the second summer following the initial cross, selection was imposed on the $F_{4:5}$ lines. For the Closed population, $F_{4:5}$ lines were planted in unreplicated single row plots at a single location and visually selected based on general agronomic traits including maturity, heading date (flowering time), plant height, stem breakage, lodging and plump kernels. For the Reopened population, $F_{4:5}$ lines were planted in single row plots at two disease nurseries with two replicates and evaluated for FHB disease severity. For lines selected with low levels of disease, harvested grain from each plot was analyzed for DON produced by the pathogen. Further selection was imposed for low DON concentration in the grain. In both the Closed and Reopened populations a ~10% selection was imposed. The selected lines were advanced to preliminary yield trials in the third summer following the cross. More detailed information on the experimental population can be found in Smith *et al.* (2010) and Smith *et al.*, (2013).

3.2.2 DNA Extraction and Genotyping

DNA was extracted from a single $F_{4:6}$ seedling from each breeding line and from a bulk of five or more seedlings for each of the exotic donor lines using the CTAB and chloroform method (Sambrook *et al.*, 1989). The 1536 SNPs assayed here (barley oligonucleotide pool assay 1 (BOPA1)) were identified based on Sanger resequencing of expressed sequence tags where Morex (an important historical Minnesota cultivar), is the most frequently represented genotype (Close *et al.*, 2009). Genotyping was conducted at the USDA-ARS Regional Small Grains Genotyping Laboratory at Fargo, ND. All lines

were genotyped with BOPA1 SNPs using Illumina GoldenGate technology (Illumina, San Diego, CA). Genotypes were called using the Illumina Beadstation software. Eleven of the 13 donor lines were genotyped. Genotype and pedigree data from the Closed and Reopened panels are available in The Triticeae Toolbox (http://triticeaetoolbox.org/), an updated version of The Hordeum Toolbox (Blake *et al.*, 2012). Genotypic and phenotypic data for individual lines is available for download by selecting populations labeled "MN Reopened" and "MN Closed" under "select lines by properties".

3.2.3 Data Analysis

As a part of SNP data quality control, all SNPs monomorphic in the combined Closed and Reopened panels were removed. We also removed SNPs and individual samples with ≥10% missing data or with ≥10% observed heterozygosity. Barley is a selfing species and progeny from breeding crosses were genotyped at the $F_{4:6}$ generation, so SNPs or samples with elevated heterozygosity are likely due to genotyping errors. SNP positions were based on the consensus genetic map of Muñoz-Amatriaín *et al.* (2011) and are depicted in Figure S3.1.

SNPs were annotated to determine the genes of origin. Annotations were performed using the SNP annotation tool, SNPMeta (Kono *et al.*, 2013) based on the contextual sequence used for the Illumina SNP assay design (Close *et al.*, 2009). SNP contextual sequences were used as BLAST queries against NCBI's nucleotide (nt) database. The best BLAST hit with an annotated coding sequence was downloaded and aligned to the SNP contextual sequence. Information including gene name, and whether the SNP causes

a synonymous or nonsynonymous change was recorded for each SNP. The majority of annotations originate from a large collection of full-length cDNAs (Sato *et al.*, 2009a; Matsumoto *et al.*, 2011). The number of annotated barley genes within genomic regions identified in our studies was inferred using relative genetic map positions from the barley GenomeZipper (Mayer *et al.*, 2011).

Linkage disequilibrium (LD) measured as r^2 (correlation coefficient) (Hill and Robertson, 1968) for all possible pairwise comparisons on each linkage group was calculated in R (R Development Core Team, 2011), and the R package LDheatmap (Shin *et al.*, 2006) was used to generate plots of LD relative to genetic distance. The package hierfstat (Goudet, 2005) was used to calculate haploid F_{ST} for each SNP based on comparison of the Closed and Reopened panel with heterozygous SNPs treated as missing data. An empirical threshold of the top 2.5% of F_{ST} values on a per SNP basis was used to identify F_{ST} values that differed dramatically from the genome-wide average. The R package ape (Paradis *et al.*, 2004) was used to calculate percent pairwise difference between the Closed or Reopened panel and donor lines. Other SNP descriptive statistics were calculated using the programs compute and sharedPoly from the libsequence C++ library (Thornton, 2003), including number of segregating sites, number of singletons, and mean per-SNP pairwise diversity (Tajima, 1983) within each of the Ancestral, Closed and Reopened panel and number of private and shared SNPs.

Segments of identity by state (IBS) were identified using GERMLINE (Gusev *et al.*, 2009). Each SNP and a minimum of five adjacent SNPs were considered sequentially, with length of shared haplotypes extended until mismatch. IBS was calculated based on

comparison of each of the Reopened lines and their respective donor or donors (Table S3.1). On each linkage group, we jointly considered all IBS segments based on each donor line, thus there were many overlapping IBS segments along the linkage group for each donor line. The number of IBS segment at each SNP was determined by summing over the number of lines in the Reopened panel that included an IBS segment for each SNP.

3.2.4 Simulation

To determine if patterns of allele frequency differentiation between the Closed and Reopened pattern could occur in the absence of selection, we performed coalescent simulation implemented in the program ms (Hudson, 2002) (see details in Supporting Information). An initial set of simulations was focused on differentiation between the ancestral population (donor lines) and the MN breeding population, which forms the basis of the Closed panel (Figure S3.2). The Ancestral and Closed panel were each represented by 120 chromosomes. The Ancestral panel includes the 11 donor parents genotyped in this study supplemented with 109 lines chosen at random from parents for the barley nested association mapping (NAM) population to balance the panel size to the same as the Closed and Reopened panels and better represent the donor population. The NAM parents are a randomly chosen sample of USDA National Small Grains Collection and thus represent the diverse panel of cultivated barley lines serving as a source for the donor population.

The folded site frequency spectrum (SFS) for the genotyped SNPs is skewed toward common variants (Close *et al*., 2009) (Figure S3.3A). To simulate ascertainment bias, we used a custom Python script that conditions on a discovery panel of fixed size, and a minimum minor allele frequency for samples in the discovery panel. The first *n* chromosomes in the simulation are designated as the discovery panel, and sites that have a minor allele frequency below a user-defined threshold in the discovery panel are removed from the simulation dataset. If migration matrices are specified, then the script assumes that the first population listed is the population in which discovery is performed. Thus SFS in simulation reflects what is observed in the empirical data,

The mutation parameter $\theta = 4N_0\mu$ and crossover rate parameter $\rho = 4N_0r$, where N_0 is the effective population size of the ancestral population, were adjusted to reflect observed values of percent pairwise diversity and levels of LD (mean r^2) calculated using tools from the libsequence library (Thornton, 2003) for each linkage group. We simulated a bottleneck in the establishment of the Closed panel that started at time T_1 and ended at T_2 (Figure S3.2). Time T_1 was set to 8000 generations before present, when barley began to be disseminated from Western Asia (Pinhasi *et al.*, 2005; Pinhasi *et al.*, 2012), and T_2 varied over a uniform distribution of 15 to 8000 generations U(15,8000). The reopened panel started ~15 generations ago (T_3). The relative size of the Closed panel is a proportion of the donor population U(0,0.02). We refined this uniform interval for the relative size based on initial simulations.

To determine the likely time of the end of the bottleneck (T_2) and the relative size of the Closed panel, we performed one million simulations and compared the simulated and

observed values of pairwise diversity (P_S and P_O) in the Closed panel. Using rejection sampling, we retained simulations if $|P_S - P_O|/P_O < \varepsilon$. After preliminary survey, we chose the acceptance rate of $\varepsilon = 0.2$ and confirmed that any choice of ε did not affect the result (not shown). The end of the bottleneck (T_2) and the relative size were determined by averaging these simulations.

To determine the expected distribution for F_{ST} values between the Closed panel and the Reopened panel based on a neutral demographic scenario involving only migration and introgression, we compared P_S and P_O in the Reopened panel to estimate the migration rate from the donor lines to the Reopened panel. The prior distribution of the migration rate was sampled from U(0,10000). The migration rate is based on $4N_0m$. We retained simulations if $|P_S - P_O|/P_O < \varepsilon$. The most likely migration rate was determined by averaging these simulations. Using the most likely migration rate from the donor lines to the Reopened panel, we simulated the complete population history and calculated F_{ST} between the simulated Closed and Reopened panels using 100,000 simulations.

3.3 Results

3.3.1 Summary statistics for the Closed and Reopened panels

Quality control resulted in a data set with 990 SNPs in 237 lines. This included the elimination of 546 SNPs, 465 of which were monomorphic in both panels. We also eliminated two samples in the Closed panel and one sample in the Reopened panel due to SNP genotype quality. The average observed heterozygosity across the complete data set was 0.43%. There were 54 SNPs private to the Closed panel versus 482 SNPs private to the Reopened panel (Table 3.1). There were 478 SNPs and 895 SNPs that were shared between the donor lines and the Closed and Reopened panels respectively (Table 3.1). Pairwise diversity in the Closed panel was 0.15 genome-wide versus 0.16 in the Reopened panel, indicating greater similarity among lines in the Closed panel (Table 3.1).

3.3.2 Allele frequency differences between the Closed and Reopened panels

Genome-wide F_{ST} averaged 0.057 between the Closed and Reopened panels. Three genomic regions with F_{ST} values exceeding the 97.5th percentile ($F_{ST} \geq 0.315$) were identified on linkage groups 2H, 4H, and 6H (Figure 3.2). Also, a single SNP on 5H exceeded the F_{ST} threshold. In the high F_{ST} regions on 2H and 4H the majority of SNPs (nine out of 11 SNPs on 2H and eight out of nine SNPs on 4H) were above the 97.5th

Table 3.1 Summary statistics for lines in the Closed and Reopened panels.
The summary statistics reported include sample size (n), # of segregating sites (S), # of singletons (s), average F_{IS}, mean pairwise diversity scaled by the number of segregating sites, percent pairwise difference with donor lines, # of private SNPs in each population (S_p) and # of shared SNPs in each comparison (S_{sh}). SD: standard deviation.

Panel	n	S	s	F_{IS}	Mean pairwise diversity	Pairwise difference with donor lines (SD)	S_p	S_{sh}	S_{sh} (with donor lines)
Ancestral	120	933	11	0.997	0.41	-	-	-	-
Closed	118	502	168	0.945	0.15	0.084 (0.163)	54	448	478
Reopened	119	930	112	0.959	0.16	0.137 (0.130)	482		895

percentile threshold (Table 3.2). On 6H, only four of 20 SNPs in >~10 cM high F_{ST} region exceeded the 97.5th percentile threshold.

Comparison of minor allele frequency (MAF) demonstrates that genome-wide, there were far fewer SNPs segregating in the Closed than the Reopened panel (Figure S3.1). The three primary regions on 2H, 4H, and 6H with the largest difference in MAF between the Closed and Reopened panels corresponded to the three high F_{ST} blocks on these linkage groups. In the Closed panel, all SNPs in the high F_{ST} block on 4H were monomorphic, but they were polymorphic in the Reopened panel (Figure 3.2). F_{ST} plotted relative to MAF also showed the three clusters on these three linkage groups and indicated that SNPs with high F_{ST} on each linkage group also shared similar MAF (Figure S3.4).

The majority allele in the donor lines (donor allele) tended to occur at higher frequencies in the Reopened than in the Closed panel. In the high F_{ST} region on 2H, the frequency of the majority donor allele was substantially higher in the Reopened

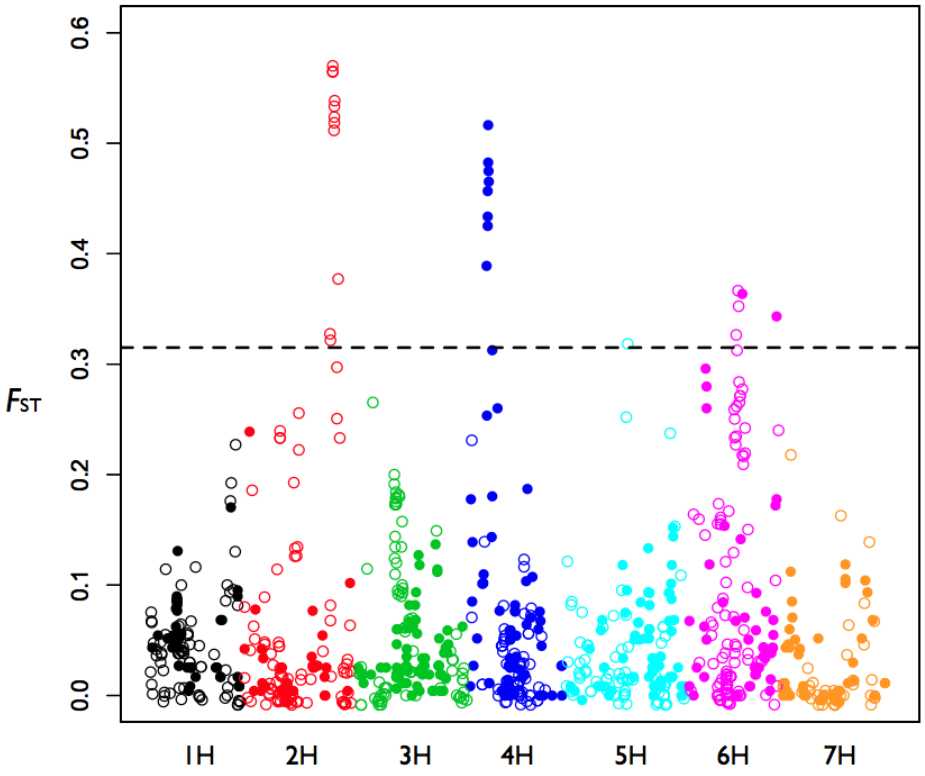


Figure 3.2 Genome-wide F_{ST} plot.
The horizontal dashed line is the 97.5th percentile. Colors correspond to linkage groups. The filled circles represent SNPs that are monomorphic in the Closed panel while open circles represent SNPs that are polymorphic in the Closed panel.

compared to the Closed panel for eight out of the 11 SNPs (Table 3.2). The region on 4H was monomorphic in the Closed panel and donor alleles introduced novel variants to the Reopened panel. All 22 SNPs in the high *F*ST regions on 6H were either monomorphic or had low MAF in the Closed panel but were more polymorphic in the Reopened panel (Table 3.2).

One feature of the MAF that could not be explained by donor introgression was a region involving 34 SNPs at 64 - 81 cM on linkage group 3H with ~0.3 MAF in the

Table 3.2 SNPs in the high F_{ST} regions on linkage groups 2H, 4H, 5H and 6H. Freq. Donor is the frequency of donor alleles (the major alleles in the donor lines). Freq. Closed and Freq. Reopened are the frequency in the Closed and the Reopened panels respectively of the donor alleles.

LG	SNP	GenBank ID	cM	F_{ST}	Freq. Donor	Freq. Closed	Freq. Reopened
2H	11_10446	XM_003560174	140.69	0.57	0.73	0.07	0.72
	11_20480	AY162186	140.69	0.57	0.91	0.08	0.72
	11_21440	AK360366	140.69	0.56	0.91	0.08	0.73
	11_21406	AK370573	142.67	0.51	0.73	0.10	0.72
	11_21370	AK362193	143.18	0.52	0.73	0.11	0.74
	11_11486	X58138	143.18	0.52	0.64	0.10	0.73
	11_21459	AM039897	143.18	0.53	0.64	0.10	0.73
	11_10109	AK362400	143.71	0.54	0.55	0.11	0.73
	11_10065	AK376660	147.37	0.30	0.64	0.59	0.15
	11_20215	AK371957	147.37	0.25	0.64	0.59	0.15
	11_20895	AK366193	149.27	0.38	0.55	0.64	0.17
4H	11_10132	AK358845	26.20	0.39	0.55	1.00	0.60
	11_20210	AK375913	26.71	0.25	0.73	1.00	0.74
	11_20422	XM_003560743	28.00	0.46	0.64	1.00	0.53
	11_21070	AK355304	28.00	0.43	0.64	1.00	0.55
	11_20302	AK372064	28.00	0.43	0.55	1.00	0.56
	11_20680	-	28.75	0.48	0.55	0.00	0.50
	11_21418	X62388	28.75	0.52	0.91	0.00	0.53
	11_20109	AK360657	29.34	0.47	0.91	0.00	0.49
	11_20777	AK364484	29.76	0.47	0.82	0.00	0.47
5H	11_21321	AK369180	97.49	0.32	0.91	0.01	0.37
6H	11_10040	AK357672	74.65	0.00	0.64	1.00	0.99
	11_10124	AK367485	74.65	0.04	0.73	0.98	0.90
	11_20015	X95863	74.65	0.33	0.91	0.03	0.42
	11_20709	AK363507	74.65	0.23	0.73	0.03	0.33
	11_20714	AK354049	75.28	0.26	0.64	0.99	0.71
	11_20468	AK362215	76.03	0.00	0.73	0.99	1.00
	11_21469	-	76.03	0.31	0.91	0.03	0.41
	11_20636	AK365203	77.53	0.37	0.91	0.01	0.41
	11_11329	AK364583	78.52	0.00	0.91	0.98	0.99
	11_20673	HQ661104	78.52	0.35	0.82	0.01	0.42
	11_20892	AK376359	79.17	0.28	0.73	0.00	0.31
	11_11349	AK361558	80.06	0.27	0.82	0.01	0.31
	11_20784	AK365537	80.06	0.00	0.82	0.99	1.00
	11_11459	AK363684	80.86	0.27	0.64	0.99	0.69

	11_21256	AK359524	80.86	0.27	0.64	0.99	0.69
	11_10469	AK355738	81.79	0.00	0.91	0.99	1.00
	11_20053	AK370710	81.79	0.14	0.64	0.00	0.15
	11_20488	AK358007	84.47	0.36	0.55	1.00	0.63
	11_20682	AY029260	84.47	0.22	0.64	0.01	0.25
	11_20969	-	84.47	0.28	0.82	0.01	0.31
	11_11111	AK366470	139.09	0.34	0.55	1.00	0.63
	11_20687	AK366288	139.09	0.18	0.73	1.00	0.81

Closed panel (Figure S3.1). There were two primary haplotypes for these 34 SNPs (data not shown).

3.3.3 Simulation

A discovery panel of eight chromosomes with a minimum minor allele count of three reflected the design parameters of the barley OPAs (Close *et al.*, 2009). In simulations of the ancestral population, this discovery scheme also closely matched the observed SFS (Figure S3.3) (paired t-test p-value = 1).

When including the Closed panel in the model, the median of simulated pairwise diversity along each linkage group in the Ancestral panel was 0.054 with 95% CI (0.042, 0.068). In the Closed panel the median of simulated pairwise diversity was 0.007 and pairwise diversity was zero in 35% of simulations. The pairwise diversity in both the Ancestral panel and the Closed panel provided a close fit to the observed data (Table S3.2). In our simulations, there was convergence in posterior density of relative size of the Closed panel, which was ~1% of the ancestral population (Figure S3.5). There was a wide interval for the most likely timing of the end of the bottleneck, which varied across the range of prior values. When we plotted the density of these two parameters together, the most likely timing of the end of the bottleneck corresponded to the highest likelihood

of the relative size, which was at 0.0015, ~900 generations ago (Figure S3.6). The relative size of the bottleneck and duration of the bottleneck were confounded as suggested by previous study (Eyre-Walker *et al.*, 1998).

Adding the Reopened panel to the simulation, and simulating migration, the estimated migration rate was $4N_0m = 4000$, which corresponded to 0.01 migrants per generation over 15 generations (Figure S3.7). In these simulations, where demography alone impacted allele frequency (i.e., where we were testing a neutral null hypothesis), the 97.5th percentile of $F_{ST} = 0.09$ between the simulated Closed and Reopened panel, versus $F_{ST} = 0.315$ in the empirical data (Figure 3.3). This suggests that in the absence of selection, demography alone is unlikely to produce the extreme values of F_{ST} observed between the Closed and Reopened panels.

3.3.4 Segments of IBS

Individual donors contributed to an average of 12 progeny in the Reopened panel. Zhedar1 contributed to the largest number of progeny, 49, while Comp351 and BT463 contributed to only a single individual (Table S3.1). The highest degree of IBS between the donor lines and their progeny in the Reopened panel on 2H and 6H overlapped the high F_{ST} regions (Figure 3.4; Figure S3.8). However, the highest IBS region on 4H did not overlap with the high F_{ST} region but rather occurred ~40 cM away (Figure 3.4). This resulted from a localized contribution of high IBS from donor Hor211 to 24 progeny; excluding this donor, the highest degree of IBS on 4H also overlapped with the high F_{ST} region for most donor lines (Figure S3.9). The degree of IBS drops dramatically at both

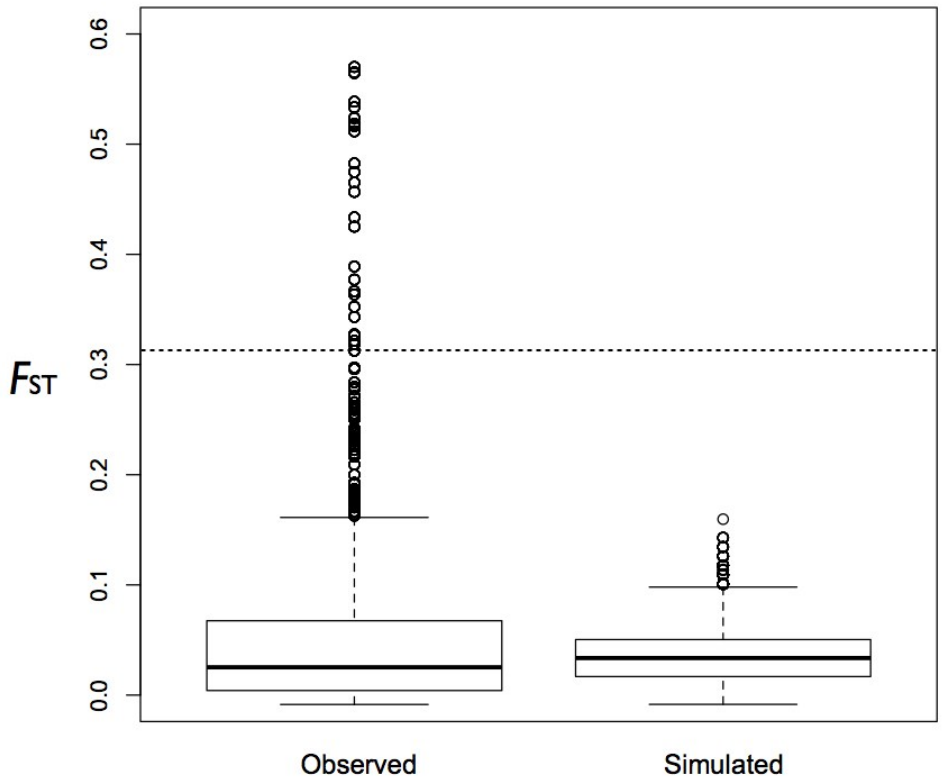


Figure 3.3 The boxplots for observed and simulated *F*ST between the Closed and Reopened panels.
Simulations were based on demography alone. The dashed horizontal line is the 97.5th percentile of *F*ST in the empirical data.

ends of the chromosome, where SNP number limits the potential to identify long segments that were IBS. Donor line Zhedar1 contributed most to the Reopened panel in the high *F*ST region on 2H (Figure 3.5) while PFC88209 contributed most to the high *F*ST regions on 4H and 6H (Figure S3.9). We summed the number of IBS segments at each SNP across the genome and across the three major high *F*ST regions. We found little correlation between the timing of introgression of donor lines and the number of IBS segments genome-wide (r^2 = 0.20) as well as in the high *F*ST regions (r^2 = 0.10).

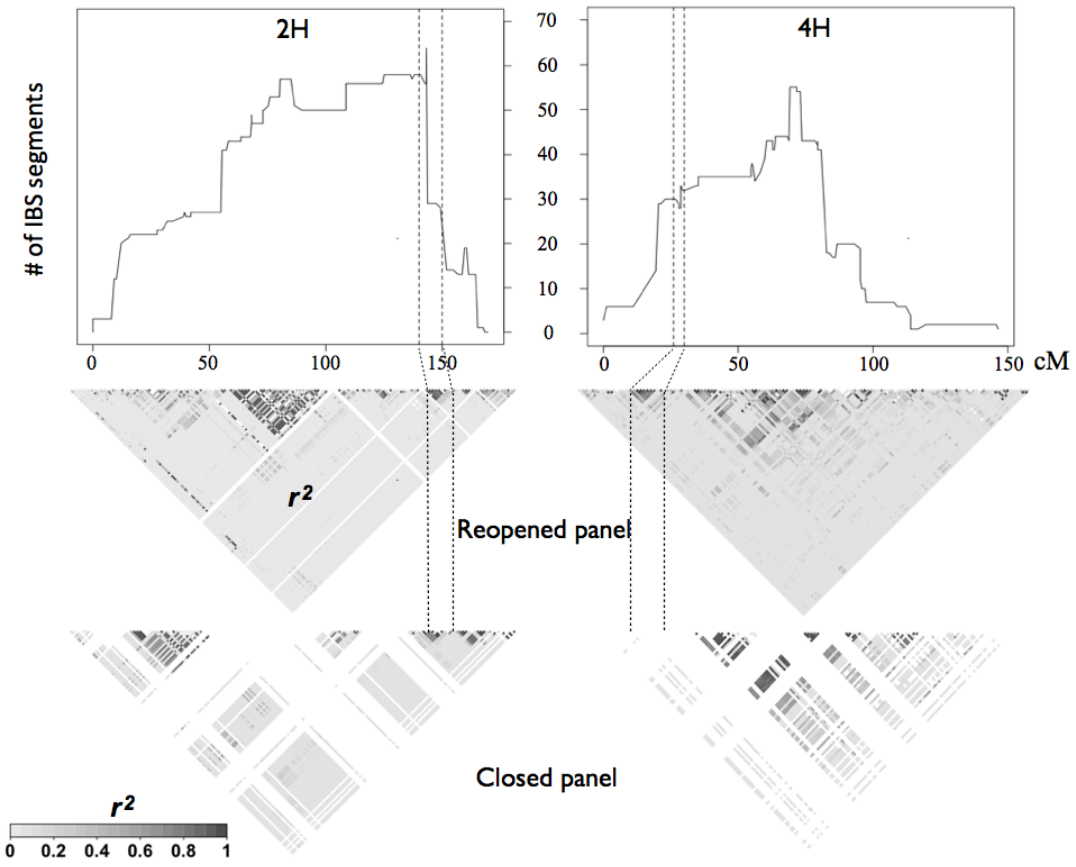


Figure 3.4 IBS and LD plot on linkage groups 2H and 4H.
The upper panel shows number of IBS segments between the donor lines and their progeny in the Reopened panel. The vertical dashed lines delimit the high *F*ST block. The middle and lower panels are LD heatmaps of the Reopened and Closed panels respectively.

3.3.5 LD in the Closed and Reopened panels

Average genome-wide LD (r^2) among all pairs of SNPs was higher in the Closed panel (0.051) than in the Reopened panel (0.028). LD between adjacent SNPs was also

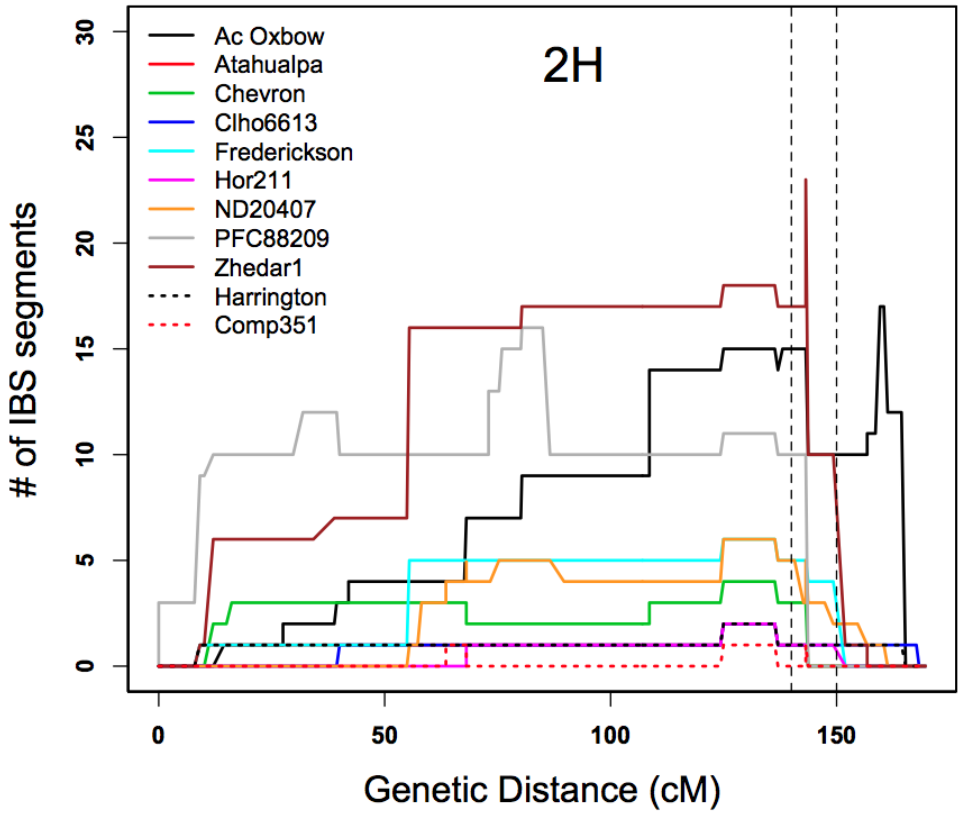


Figure 3.5 IBS between each of the donor lines and their respective progeny in the Reopened panel on 2H.
The vertical dashed lines delimit the high *F*ST block.

higher in the Closed panel (0.653) compared to the Reopened panel (0.490) (Figure S3.10).

Blocks of LD were defined as sets of at least three adjacent SNPs that showed greater LD than the median r^2 of adjacent SNPs (0.58). There were 28 blocks in the Closed panel covering a total of 65.13 cM and 45 blocks in the Reopened panel covering a total of 80.63 cM. The average block size in the Closed panel was 2.33 cM, which was

greater than that in the Reopened panel (1.79 cM). In the Closed panel, 15.9% of SNPs were in LD blocks, while 21.8% of SNPs were in blocks in the Reopened panel.

All the SNPs in the two high F_{ST} blocks on linkage groups 2H and 4H were also in high LD ($r^2 > 0.21$, the 97.5th percentile threshold) with each other in the Reopened panel (Figure 3.4). The LD pattern was less clear in the linkage group 2H block in the Closed panel ($r^2 = 0.477$ versus 0.613) and the SNPs were monomorphic in the linkage group 4H block in the Closed panel (Figure 3.4). The LD in the high F_{ST} region on 6H is similar in the Closed and Reopened panels ($r^2 = 0.438$) (Figure S3.8).

3.3.6 Comparison to previous studies

We identified markers from previous studies that occur in genetic map locations adjacent to high F_{ST} genomic regions in our comparison (see Table S3.3). The high F_{ST} regions on 2H, 4H and 6H overlapped with the relative genetic map positions of markers associated with DON concentration and FHB resistance in previous QTL mapping studies (Ma *et al.*, 2000; Mesfin *et al.*, 2003) as well as in a recent GWAS study that included elite breeding lines from four Midwest breeding programs including the University of Minnesota program (Massman *et al.*, 2011).

In addition to SNPs surveyed here (BOPA1), three additional sets of SNPs have been mapped in barley genetic mapping populations (Rostoks *et al.*, 2005; Close *et al.*, 2009; Muñoz-Amatriaín *et al.*, 2011). All BOPA1, BOPA2, pilot oligonucleotide pool assays (POPA) and Scottish Crop Research Institute (SCRI; the SCRI is now known as the James Hutton Institute). SNPs (http://bioinf.hutton.ac.uk/iselect/app/) that fall within

annotated genes in the high F_{ST} blocks on 2H, 4H and 6H are listed in Table S3.4. The genomic regions identified are ~5 cM (4H) and ~10 cM (2H and 6H) and include a minimum of 40 (on 4H) or 100 genes (on 2H and 6H).

3.4 Discussion

Several genomic regions putatively subject to selection have been identified using comparative population genetic methods in this barley experimental population. These regions show strong allele frequency differentiation between the Closed and Reopened panels, excess IBS between the Reopened panel and donor lines, and elevated LD in the Reopened panel.

3.4.1 Variability of allele frequency

In the donor lines Chevron and Frederickson, QTL contributing to FHB resistance have been mapped to the same interval found to have high F_{ST} on 2H and 6H (de la Pena *et al.*, 1999; Ma *et al.*, 2000; Mesfin *et al.*, 2003; Massman *et al.*, 2011). The SNPs in the two intervals on linkage groups 2H and 4H have F_{ST} values in the top 2.5% genome-wide. The plot of F_{ST} versus MAF (Figure S3.4) shows clusters of high F_{ST} SNPs on the same linkage group having similar MAF, which suggests these SNPs co-occur, potentially due to selection favoring a relatively small number of haplotypes. Interestingly, all SNPs in the high F_{ST} block on 2H are polymorphic within the Closed

panel but SNPs in the high F_{ST} block on 4H are monomorphic in the Closed panel (Figure 3.2; Table 3.2). Within the limits of the experiment, this suggests that selection at 4H is more likely to have acted on newly introgressed allelic variation.

In this experimental population, allele frequency changes appear to provide an effective means of identifying genomic regions subject to selection. However, there are limitations to this approach. First, selection on FHB phenotypes is at least partially confounded with selection for agronomically adaptive phenotypes, particularly in the parent used for later generation crosses in the recipient (Reopened) population. Although there are QTL associated with heading date and plant height that are coincident with FHB or DON in the high F_{ST} region on 2H and 4H, the QTL associated with heading date and plant height on 6H are not within the high F_{ST} regions based on the association mapping study in Massman *et al*., 2011. Second, the family structure within the population may tend to inflate differences in SNP frequency. Third, as with any outlier-based approach, extreme values of F_{ST} could result from stochastic processes and thus would represent false positives when attributed to selection. Finally, in the present study genetic resolution is limited and differences in allele frequency likely reflect only the general proximity of causative mutations. The level of genetic resolution provided by the high F_{ST} regions is comparable to association mapping studies in plant breeding populations (Cockram *et al.*, 2010; Massman *et al.*, 2011) and is not a major impediment to the utilization of the identified genomic regions in marker assisted breeding or genomic selection approaches.

3.4.2 Variability of IBS

Along with assaying allele frequency changes, we used IBS analysis to identify regions that were putatively subject to selection. The high IBS regions on 2H and 6H were within or adjacent to the high F_{ST} regions (Figure 3.4; Figure S3.8). The IBS analysis identified a narrower region, as the number of IBS segments at SNP 11_21459 on 2H is much higher than that at both adjacent SNPs. However, the excess IBS region in the high F_{ST} region on 4H did not have the highest number of IBS segment on this linkage group (Figure 3.4), which was primarily contributed by one donor line, Hor211 (Figure S3.9).

The IBS analysis also has limitations. The timing of introgression of donor lines could influence IBS results, as the lines introgressed recently would have more IBS segments. However, our result shows a strong correlation does not exist between the timing of introgression and the number of IBS segments, and the two donor lines PFC88209 and Zhedar1 that contributed most to the high F_{ST} regions were not among the most recently employed donors (Figure 3.1).

3.4.3 Variability of LD

The distribution and pattern of LD (r^2) differs dramatically between these two panels (Figure 3.4; Figure S3.10). The introduction of allelic diversity to the Reopened panel resulted in a larger number of polymorphic SNPs, and a reduction in both average genome-wide and average adjacent SNP LD within this population. Although average LD is lower in the Reopened panel, the LD is higher in the two high F_{ST} regions on linkage

groups 2H and 4H in the Reopened panel, as can occur in genomic regions subject to recent strong selection (Sabeti *et al.*, 2002; McVean, 2007) (Figure 3.4). Therefore, the blocks of LD in the Reopened panel on linkage groups 2H, 4H and 6H likely result from selection on haplotypes for FHB resistance or reduction in DON accumulation. A number of other factors, however, can potentially contribute to localized elevation of LD, including recent admixture (Pfaff *et al.*, 2001) or suppressed recombination due to chromosomal structural variation (Graubard, 1932; Fang *et al.*, 2012).

Introducing genetic diversity and shifting selection pressure changes the distribution and pattern of LD among markers, and therefore between markers and QTL. This suggests that it will be important to use panels of germplasm that are contemporary and relevant to current breeding goals for association analysis and generating prediction models for genomic selection. As genetic distance between two breeding populations increases, the correlation between closely linked markers in the two populations decreases. Based on this, Hamblin *et al.* (2010) cautioned against pooling data from different breeding programs for association analyses. However, similar concerns may apply to populations within a breeding program that have different breeding histories. Recent work evaluating genomic selection prediction accuracy has shown that using a training population from distinct, but closely related breeding programs, provides less accurate predictions than from a training population representing the target breeding population (Lorenz *et al.*, 2012).

3.4.4 Summary

We have employed population genetic approaches to identify genomic regions putatively subject to selection subsequent to introgression in a barley breeding experiment. The progenitor-derivative relationship between the two populations in our study is among the simplest possible scenarios for detecting the effects of recent strong selection (Innan and Kim, 2008). Other comparative analyses of plant breeding history generally deal with more diverse breeding histories over longer periods of time (Sim *et al.*, 2010; van Heerwaarden *et al.*, 2012), making inference of the selective pressure on outlier loci more difficult. The identification of the genomic regions previously associated with FHB resistance and DON concentration suggests that the comparative approach applied here is complementary to the identification of trait-associated markers through QTL and association mapping approaches.

There are several advantages to the comparative approach as applied here. The first is the relative speed and minimal expense associated with identification of putatively trait-associated loci. This experiment was conducted within the confines of a breeding program, obviating the need for multiple QTL mapping populations to identify sources of resistance. Indeed, the program developed two high-yielding malting cultivars from material represented in this population. Rasmusson is representative of the Closed population (Smith *et al.*, 2010) while Quest (Smith *et al.*, 2013), a cultivar with reduced DON accumulation is derived from donor parents Chevron and Zhedar1 and is a product of the Reopened population. Second, comparison of allele frequencies among populations will remain effective even as divergence and low minor allele frequencies within

populations minimize the potential for effective association mapping. Finally, the comparative population genetic approach is also free of *a priori* identification of phenotypes to be measured and can benefit dramatically from increased SNP density (Ross-Ibarra *et al.*, 2007; Walsh, 2008). We note that while we identified interesting genomic regions without the use of phenotype data in our two panels, the approach is not "phenotype free", but rather the result of repeated strong phenotypic selection.

The fact that we identified signals of selection for FHB resistance that are substantiated by prior mapping efforts in barley suggests that this approach may also be effective when applied to crop species without prior information from QTL mapping. Application of inexpensive genotyping to any breeding population that has a defined history of breeding and selection should provide valuable insight into the genetic architecture of the traits under selection and guidance for marker-based breeding efforts.

3.5 Supporting Information

3.5.1 Supplementary Text

The Minnesota (MN) barley breeding population (the basis of the Closed panel) has greatly reduced diversity relative to donor lines (Table 3.1). To generate simulations consistent with this difference in diversity among populations, we simulate the establishment of bottleneck associated with the establishment of the MN population. In ms simulations, we use values U(0.000025, 0.008) and U(0, 0.02) for the end of the bottleneck and relative size of the Closed panel. The population represented by the Reopened panel started ~15 generations ago. For scaling in ms, with time scaled in $4*N_0$, we use N_0 = 150,000 based on $\theta = 4N_0\mu = 0.003$ and $\mu = 5*10^{-9}$, where μ is mutation rate per site per generation. We assume one generation per year. The bottleneck started at $8000/4*N_0 = 0.013$. The Reopened panel started 15 generations before present, $15/4*N_0 = 0.000025$. The end of the bottleneck can be anytime between the start of the Reopened panel (0.000025) and the start of the bottleneck (0.013). Therefore, the relative size of the Closed panel is sampled from a uniform distribution U(0, 0.02). The end of bottleneck is also sampled from a uniform distribution U(0.000025, 0.013). Based on initial simulations, we refined the interval to be U(0.000025, 0.008) in the final simulation.

The command line for simulating the ancestral population and the Closed panel is:
ms 240 1000000 -t 150 -r 1000 1000 -I 2 120 120 -n 2 0.025 -en tbs 2 tbs -ej 0.013 2 1

The command line for simulating all populations is

ms 360 100000 -t 150 -r 1000 1000 -I 3 120 120 120 -n 2 0.025 -en 0.0015 2 0.01 -n 3 0.09 -ej 0.0133 2 1 -m 3 1 tbs -ej 0.000025 3 2 0

3.5.2 Supplementary Figures

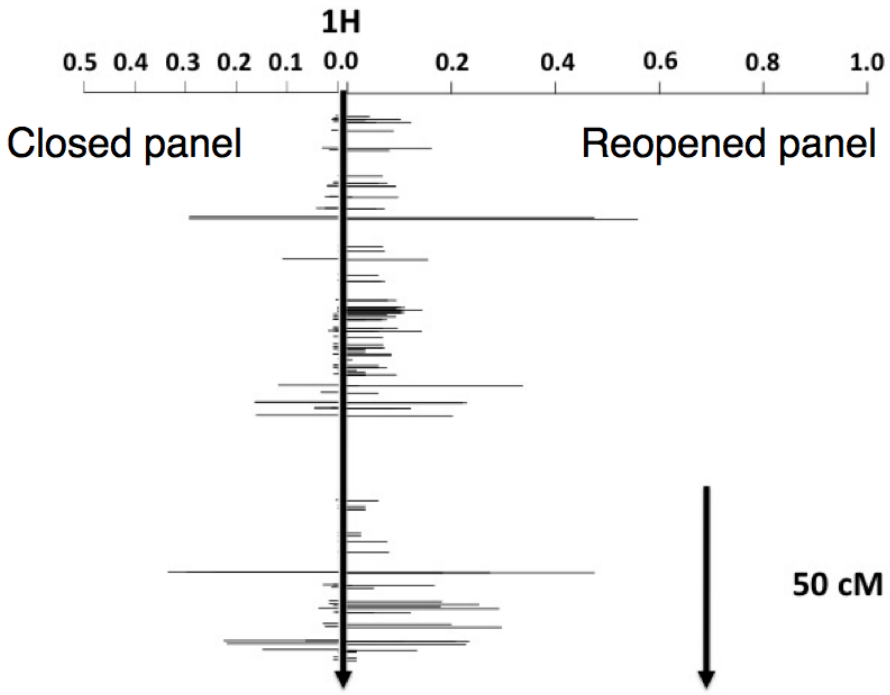


Figure S3.1 SNP positions and allele frequency comparison of the Closed and Reopened panels on each linkage group.
The frequency of the minor allele in the Closed panel is shown. The increase in frequency of some SNPs in the Reopened panel results in SNP states exceeding 50% in frequency. The red lines correspond to the SNPs in the high F_{ST} blocks.

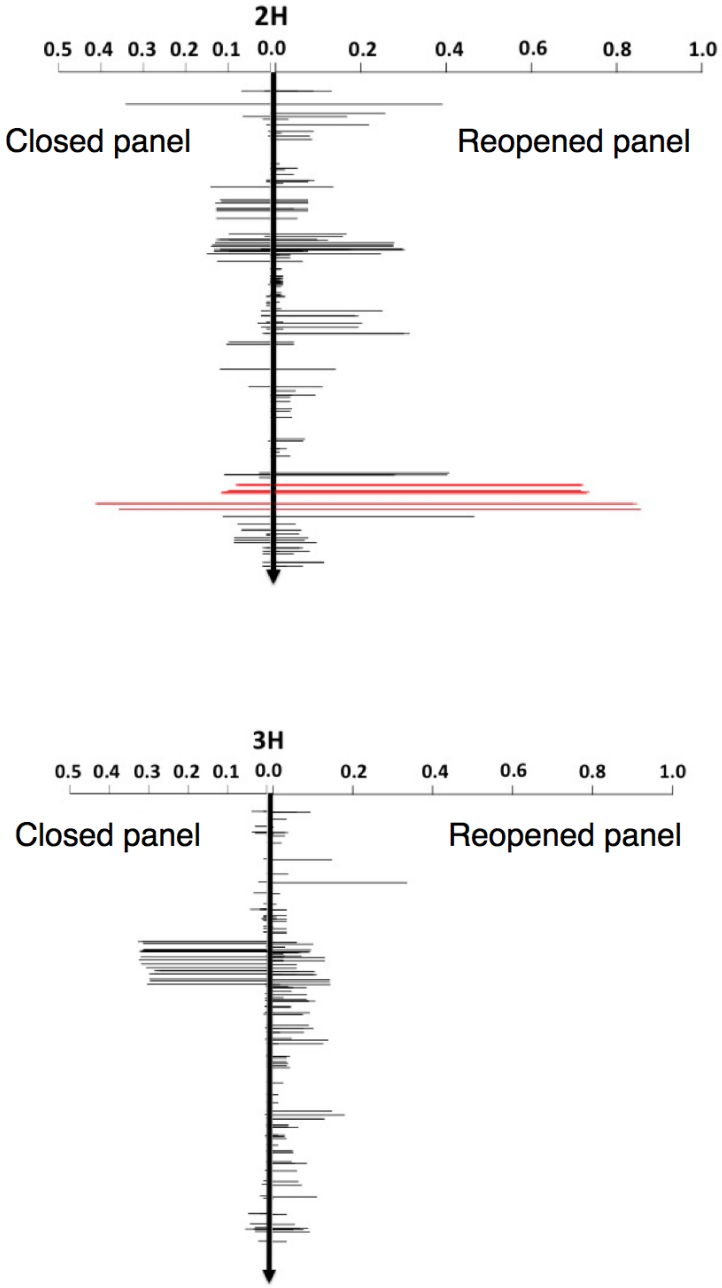


Figure S3.1 cont.

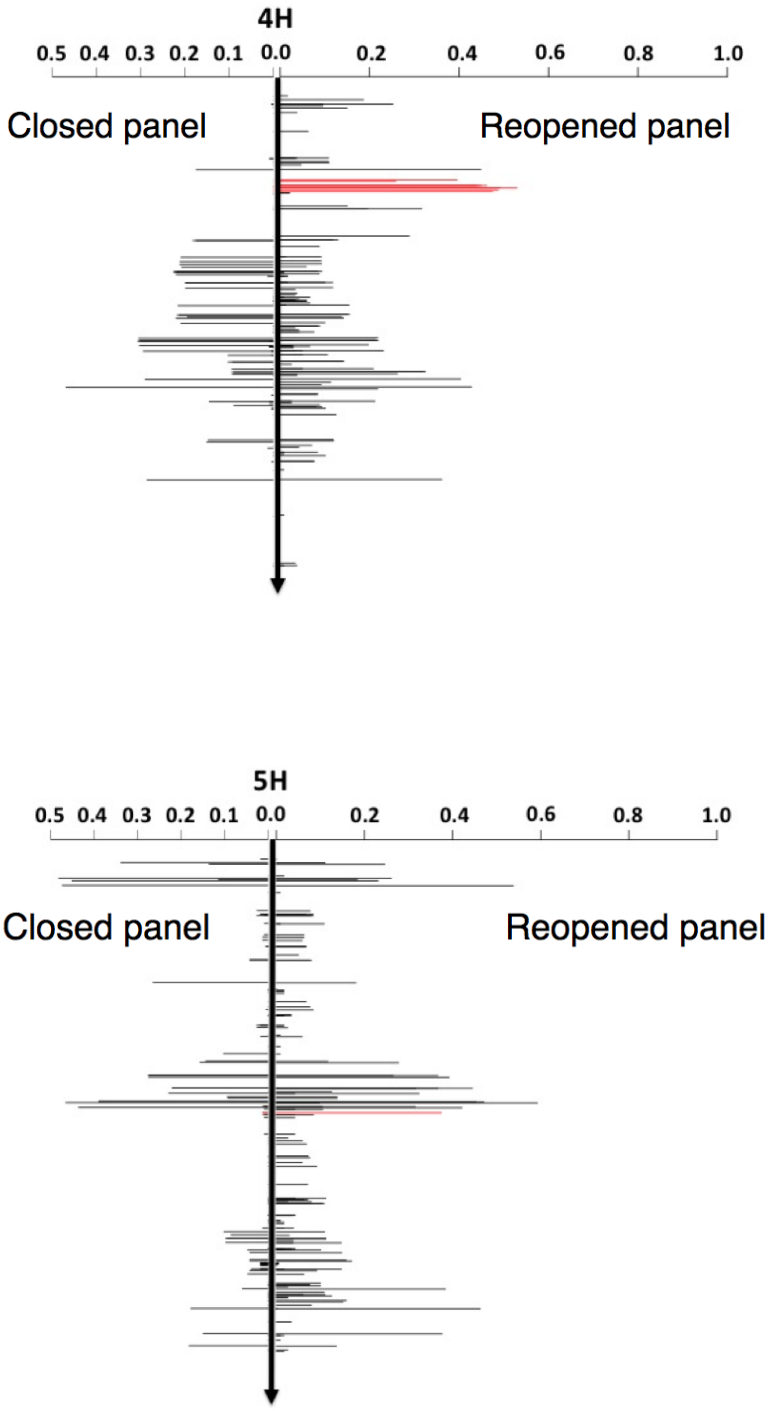


Figure S3.1 cont.

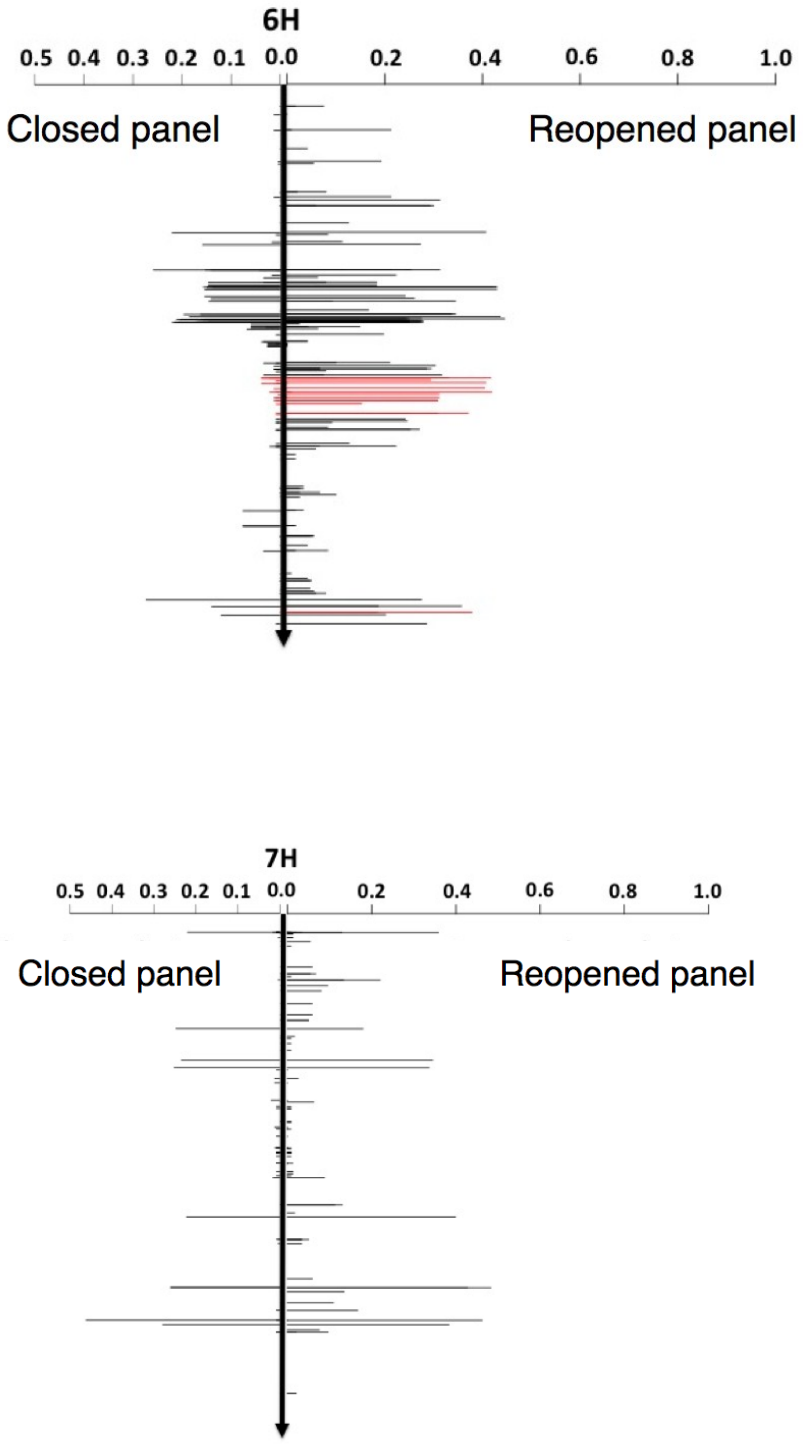


Figure S3.1 cont.

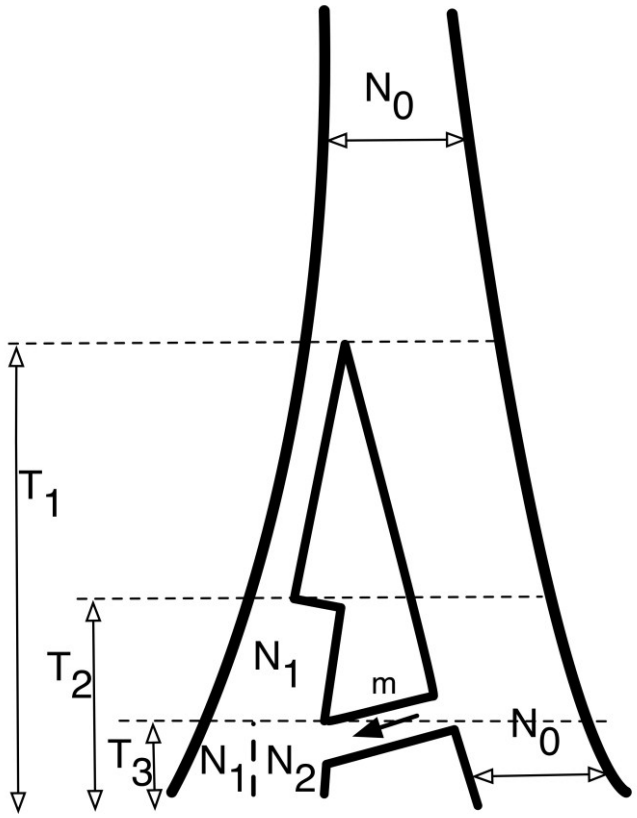


Figure S3.2 Population history
N_0, N_1 and N_2 stand for the Ancestral population, the Closed population and the Reopened population. T_1 is the start of the bottleneck population, ~8000 generations before present. T_2 is the end of the bottleneck and T_3 is the start of the Reopened population, ~15 generations before present. Migration is from the Ancestral population to the Reopened population.

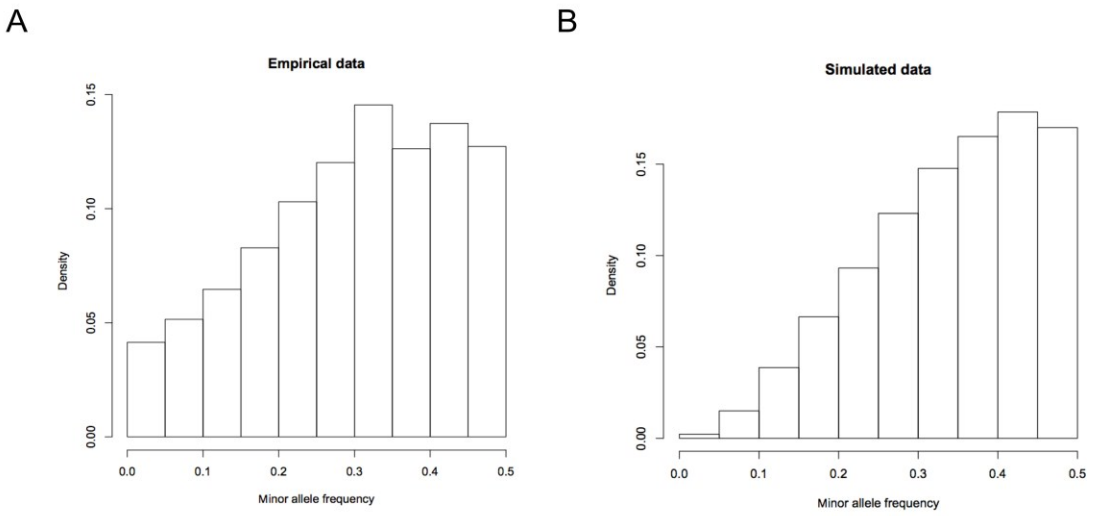


Figure S3.3 Comparison of observed and simulated SFS in the Ancestral panel
(A) Observed SFS in the Ancestral panel. (B) Simulated SFS in the Ancestral panel using a discovery panel with eight chromosomes and a minor allele count of three.

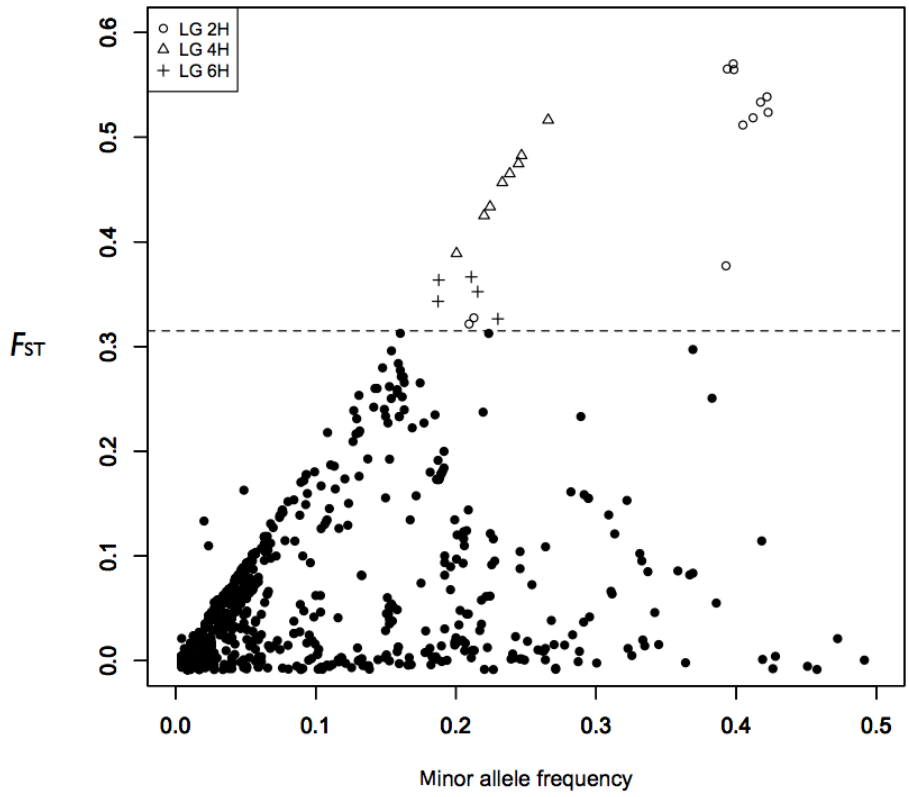


Figure S3.4 F_{ST} value versus minor allele frequency.
The horizontal dashed line corresponds to genome-wide 97.5th percentile of F_{ST} values. All SNPs below the threshold are shown as solid black points. SNPs above the threshold are shown in three symbols corresponding to each of the three linkage groups. Minor allele frequency is based on the whole dataset, including both the Closed and Reopened panels.

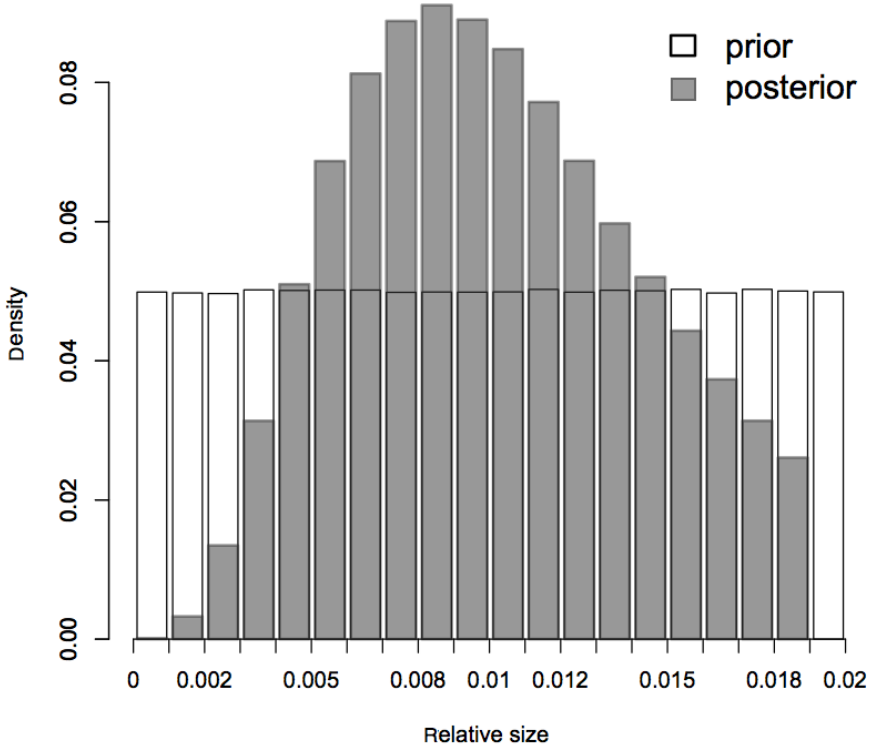


Figure S3.5 Prior and posterior density of relative size of the Closed panel from simulations.

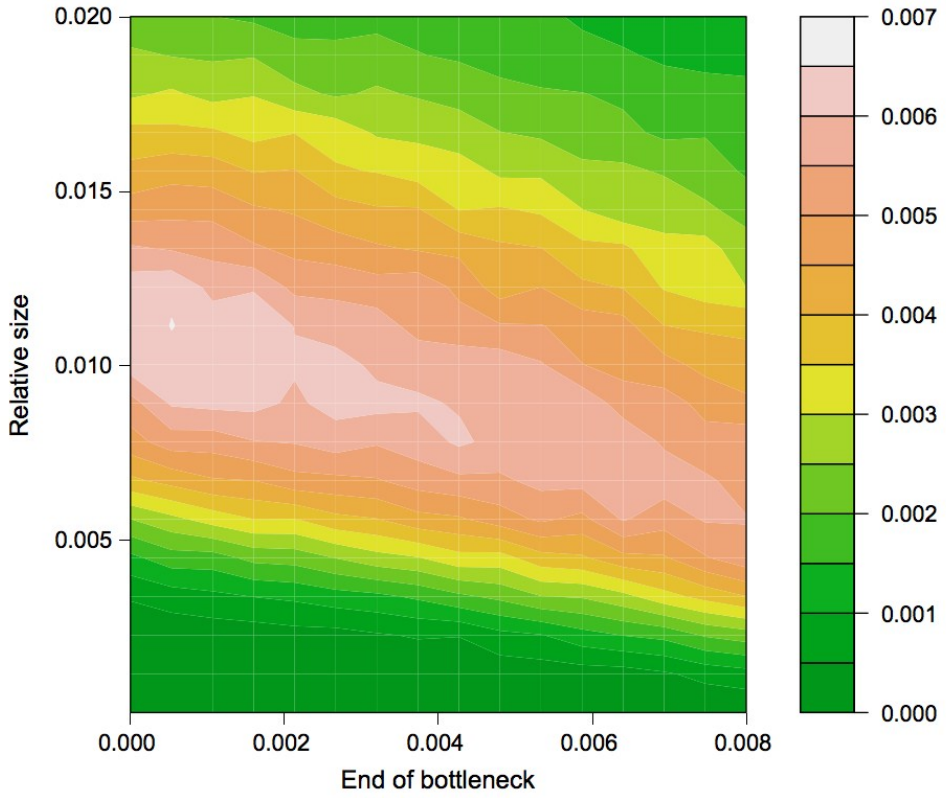


Figure S3.6 The heatmap of bottleneck.
The x-axis is the timing of the end of the bottleneck. The y-axis is the relative size of the bottleneck.

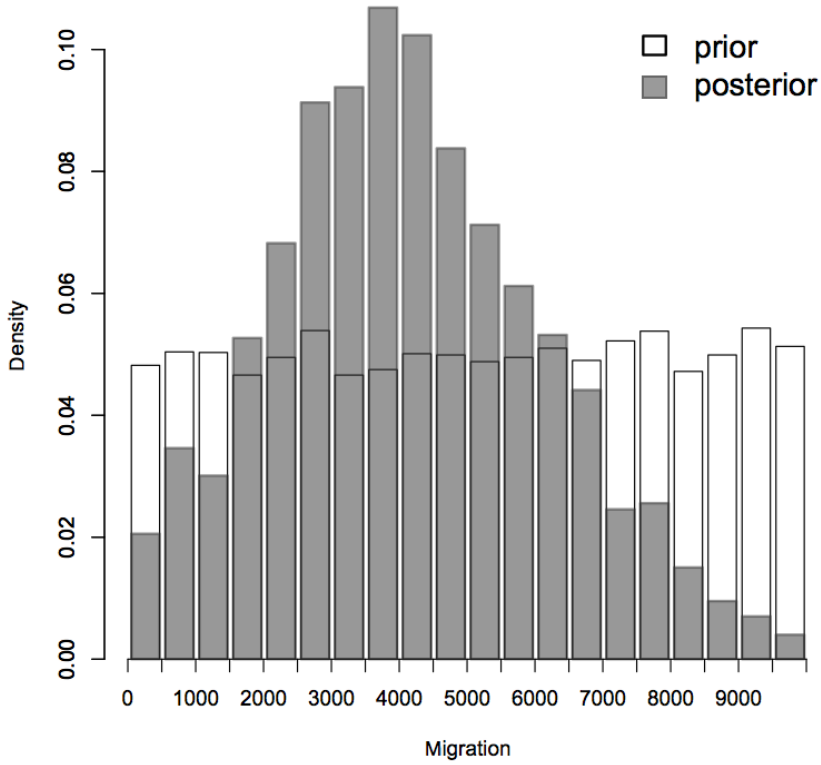


Figure S3.7 Prior and posterior density of migration rate from the Ancestral panel to the Reopened panel.

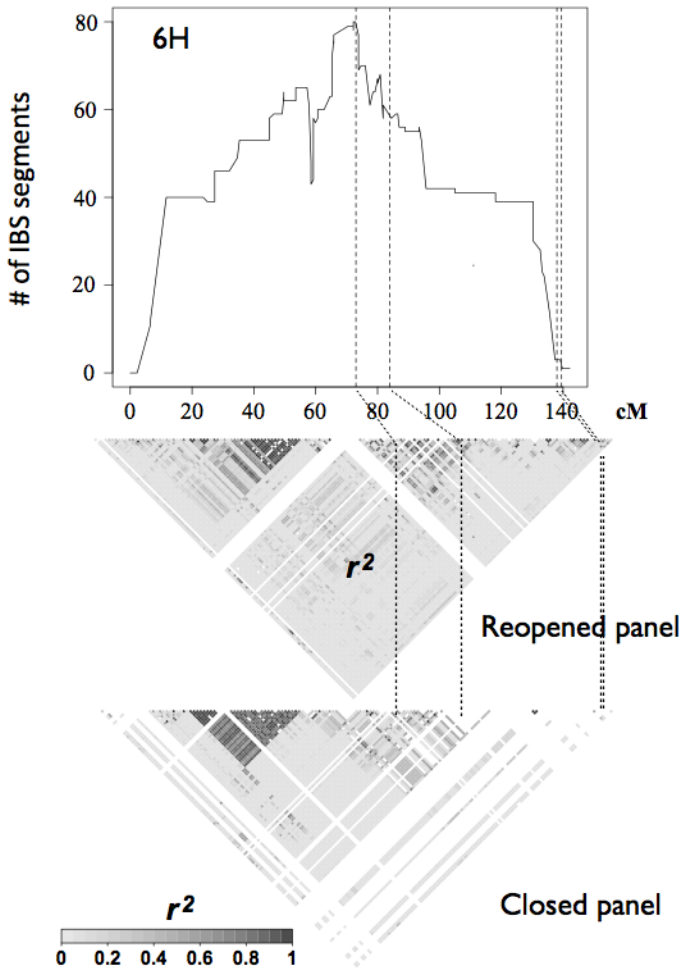
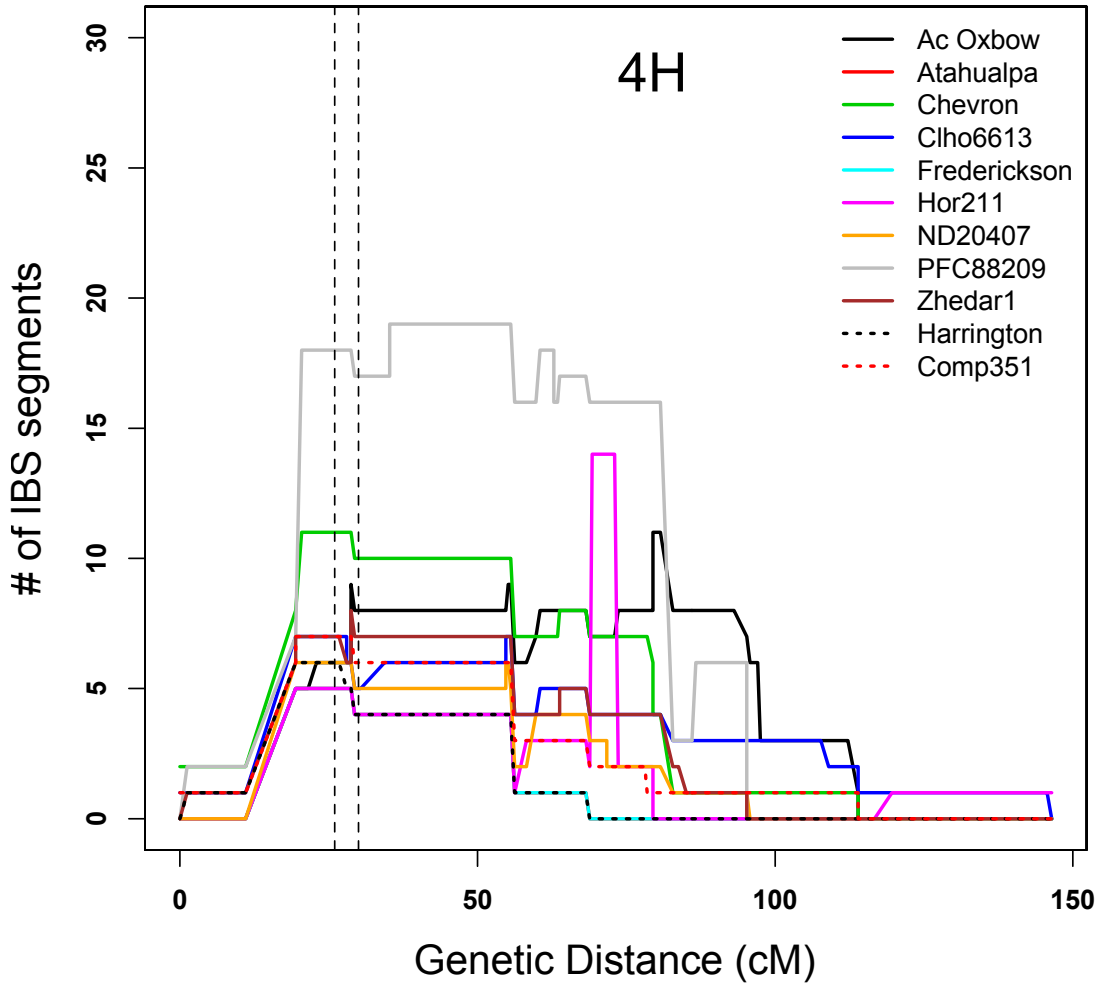


Figure S3.8 IBS and LD plot on linkage group 6H.
The upper panel shows the number of IBS segments between the donor lines and their progeny in the Reopened panel. The vertical dashed lines delimit the high F_{ST} block. The middle and lower panels are the LD heatmaps of the Reopened and Closed panels respectively.


4H
Ac Oxbow
Atahualpa
Chevron
Clho6613
Frederickson
Hor211
ND20407
PFC88209
Zhedar1
Harrington
Comp351
# of IBS segments
0
5
10
15
20
25
30
0
50
100
150
Genetic Distance (cM)

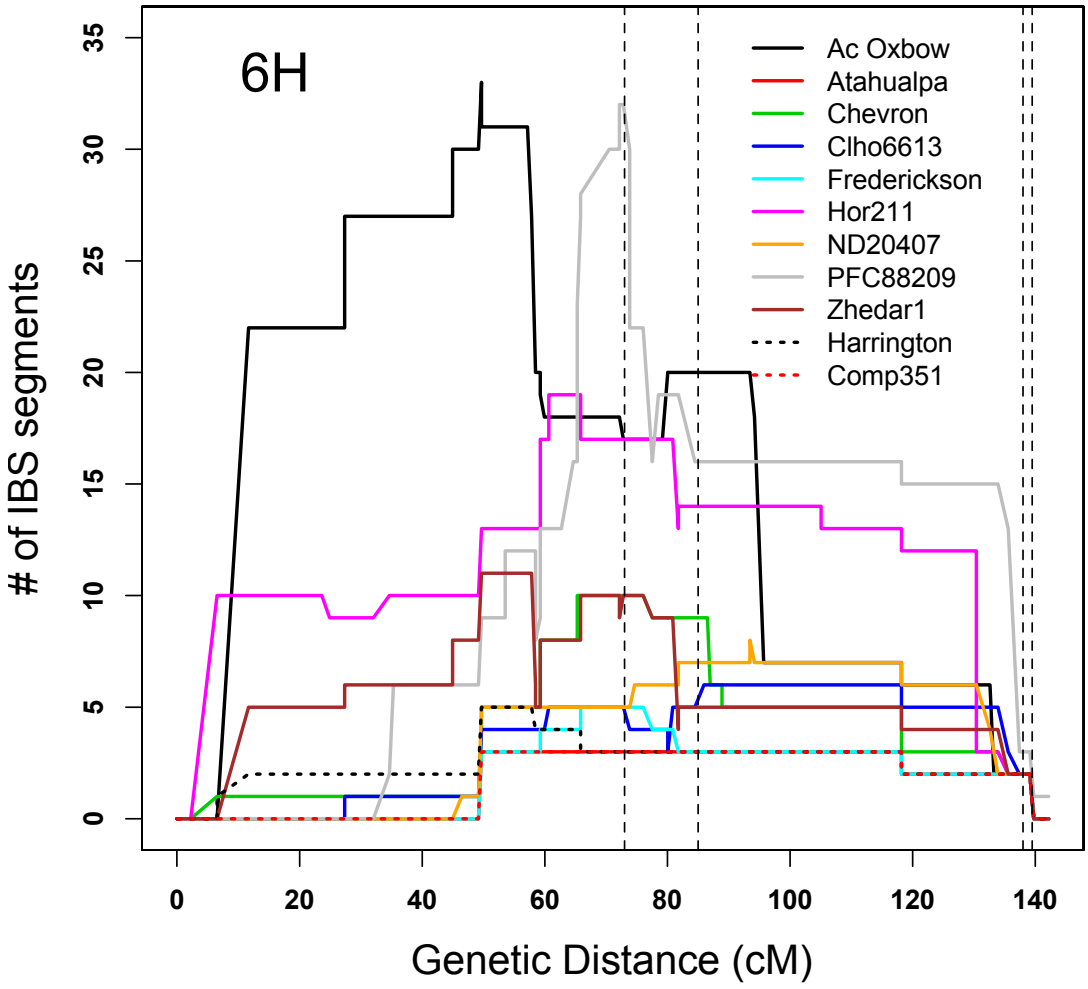


Figure S3.9 IBS between each of the donor lines and their respective progeny in the Reopened panel on 4H and 6H.
The vertical dashed lines delimit the high F_{ST} block.

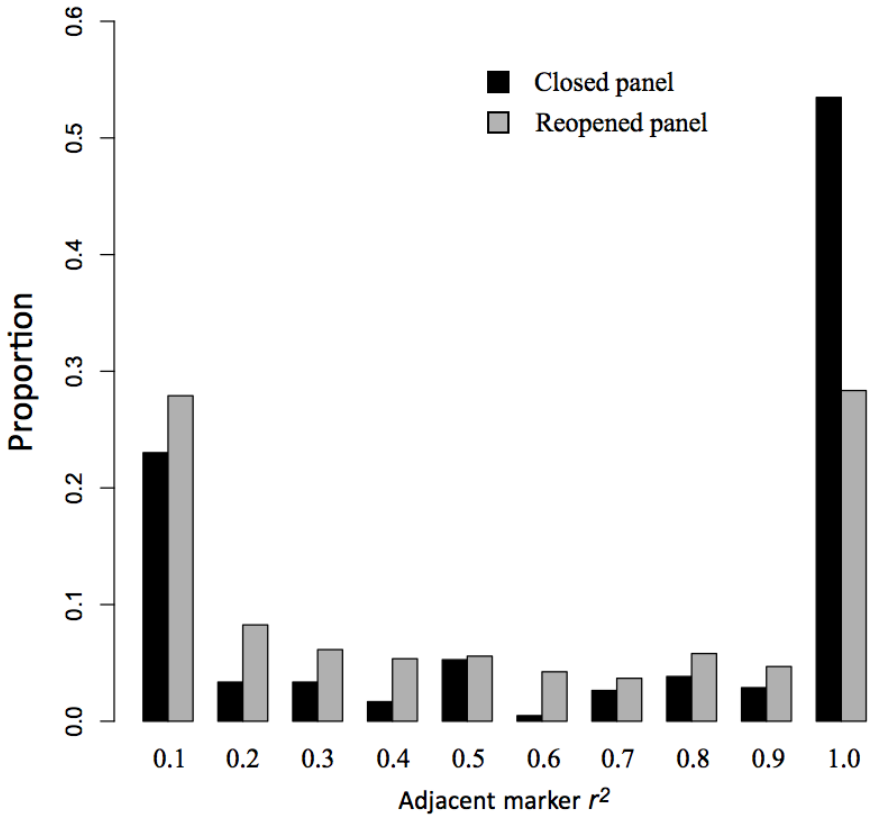


Figure S3.10 Percent of adjacent SNPs at varying levels of LD in the Closed and Reopened panel.

3.5.3 Supplementary Tables

Table S3.1 The donor line/lines of each line in the Reopened panel.

Reopened lines	Donor lines
C113.004	Chevron
C119.002	Chevron
FEG59.09	Ac Oxbow
FEG60.27	BT463
FEG61.37	Clho6613
FEG63.16	Chevron
FEG63.56	Chevron
FEG65.02	Zhedar1
FEG66.05	Zhedar1
FEG66.08	Zhedar1
FEG66.21	Zhedar1
FEG66.31	Zhedar1
FEG67.12	Frederickson
FEG69.24	PFC88209
FEG69.38	PFC88209
FEG73.13	Hor211
FEG73.49	Hor211
FEG74.18	Hor211
FEG74.19	Hor211
FEG75.39	Hor211
FEG80.06	Zhedar1
FEG80.53	Zhedar1
FEG81.58	Harrington
FEG81.60	Harrington
FEG82.16	Chevron
FEG86.03	Hor211
FEG86.53	Hor211
FEG88.73	Atahualpa, Zhedar1
FEG88.87	Atahualpa, Zhedar1
FEG89.73	Hor211

FEG90.31	Zhedar1, Atahualpa
FEG90.35	Zhedar1, Atahualpa
FEG91.28	PFC88209, Frederickson
FEG93.12	Frederickson
FEG93.36	Frederickson
FEG94.20	Zhedar1
FEG94.41	Zhedar1
FEG96.06	Ac Oxbow
FEG96.55	Ac Oxbow
FEG97.14	Ac Oxbow
FEG97.44	Ac Oxbow
FEG98.53	PFC88209
FEG99.10	Ac Oxbow
FEG99.51	Ac Oxbow
FEG100.17	Zhedar1
FEG100.33	Zhedar1
FEG100.41	Zhedar1
FEG100.47	Zhedar1
FEG103.44	Ac Oxbow, Harrington
FEG103.45	Ac Oxbow, Harrington
FEG104.63	Zhedar1
FEG104.89	Zhedar1
FEG105.33	PFC88209
FEG105.59	PFC88209
FEG109.13	Ac Oxbow
FEG109.44	Ac Oxbow
FEG109.54	Ac Oxbow
FEG111.10	Ac Oxbow, Zhedar1
FEG111.13	Ac Oxbow, Zhedar1
FEG111.24	Ac Oxbow, Zhedar1
FEG112.14	Ac Oxbow, Atahualpa
FEG113.85	Ac Oxbow, Zhedar1
FEG114.33	Clho6613
FEG116.05	Zhedar1

FEG117.24	Zhedar1
FEG118.05	PFC88209
FEG118.41	PFC88209
FEG118.69	PFC88209
FEG121.03	Zhedar1, Ac Oxbow
FEG121.16	Zhedar1, Ac Oxbow
FEG121.29	Zhedar1, Ac Oxbow
FEG121.43	Zhedar1, Ac Oxbow
FEG122.36	Hor211, PFC88209
FEG122.50	Hor211, PFC88209
FEG122.92	Hor211, PFC88209
FEG124.35	PFC88209
FEG125.46	Zhedar1
FEG125.69	Zhedar1
FEG126.08	Zhedar1
FEG126.14	Zhedar1
FEG129.41	Frederickson
FEG129.60	Frederickson
FEG132.05	Zhedar1, Frederickson
FEG132.63	Zhedar1, Frederickson
FEG138.08	Zhedar1, Hor211
FEG138.27	Zhedar1, Hor211
FEG141.18	Ac Oxbow
FEG141.20	Ac Oxbow
FEG142.13	Zhedar1, Hor211
FEG142.28	Zhedar1, Hor211
FEG142.55	Zhedar1, Hor211
FEG144.21	Ac Oxbow, Hor211
FEG144.27	Ac Oxbow, Hor211
FEG144.68	Ac Oxbow, Hor211
FEG146.09	Frederickson
FEG146.46	Frederickson
FEG146.68	Frederickson
FEG147.03	Zhedar1, Atahualpa

FEG147.14	Zhedar1, Atahualpa
FEG147.63	Zhedar1, Atahualpa
FEG148.22	Ac Oxbow
FEG148.56	Ac Oxbow
FEG149.18	ND20407
FEG149.65	ND20407
FEG150.42	ND20493
FEG150.49	ND20493
FEG153.22	Zhedar1
FEG155.07	Ac Oxbow
FEG156.09	Zhedar1
FEG161.03	Ac Oxbow
FEG162.22	Ac Oxbow
FEG163.21	Zhedar1
FEG164.33	Hor211, PFC88209
FEG166.38	Zhedar1
FEG168.09	Comp351
FEG169.47	Hor211
FEG170.07	Hor211
FEG172.40	Hor211
FEG175.57	Zhedar1, Hor211

Table S3.2 The observed pairwise diversity (scaled by the number of segregating sites) for each linkage group and the median of simulated pairwise diversity in the Ancestral panel, Closed, and Reopened panel.

LG	Ancestral panel	Closed panel	Reopened panel
1H	0.049	0.002	0.004
2H	0.060	0.004	0.004
3H	0.072	0.001	0.005
4H	0.057	0.001	0.005
5H	0.068	0.004	0.007
6H	0.059	0.007	0.011
7H	0.044	0.002	0.002
Simulated	0.054	0.007	0.014

Table S3.3 Markers from previous studies that are within or flanking (~5 cM) the high F_{ST} blocks and their estimated positions.

Linkage group	Marker	Position (cM)
2H	ABC252	141.89 – 142.42
	CDO373	
	F3hA	
	MWG5208	
	pKABA1	
	Ebmc0415	142.42 – 143.72
	Cnx1	
	BCD135	146.05 – 147.93
	Gln2	
	KG004.1	
	KG004.2	
	ABC157	148.58 – 150.55
	Zeo1	150.55 – 152.64
	HVM40	19.27 – 20.26

4H	CDO669A	20.68 – 21.33
	Ole1	24.49 – 25.23
	BCD402B	26.47
	CDO542	29.08 – 29.49
	DsT-29	
	CDO122	30.10
	BCD351D	30.10 – 31.41
	INT-C	30.37
	MWG635A	31.41 – 32.62
	BCD265B	34.92 – 37.54
	BCD808B	
5H	Scssr05939	94.25 – 94.66
6H	ksuD17	65.89 – 69.87
	G57	
	ABC163	71.52 – 72.17
	ABG379	
	Bmac0218C	73.18 – 73.98
	ABG388	74.60 – 75.34
	CDO507	
	ABC175	77.82 – 78.46
	RZ323	79.34 – 80.13
	ksuA3D	
	ABC1708	
	Scsnp21226	81.05 – 82.53
	MWG820	83.71 – 85.17
	cMWG684D	88.08 – 88.78
	MWG514	136.17 – 137.84
	MWG798A	
	ABG725	

	DAK213C	138.49 – 140.92
	DsT-71	

Table S3.4 BOPA, POPA and SCRI SNPs within genes of known function in the high F_{ST} blocks and their respective gene products.

LG	SNP	GenBankID	cM	Silent	Gene	Product
2H	SCRI_RS_173017	NM_001073041	139.9	No	Os12g0256900	hypothetic protein
	11_10446	XM_003560174	140.69	No	LOC100837523	serine carboxypeptidase-like
	11_20480	AY162186	140.69	Yes	exin1	Extracellular invertase
	SCRI_RS_15119	DQ163025	141.5	Yes	VTE5	phytol kinase
	11_21459	AM039897	143.18	No	ahh1	S-adenosyl-L-homocysteine hydrolase
	2_1484	AB058924	143.71	Yes	HvPKABA1	protein kinase HvPKABA1
	11_10656	XM_003580700	145.69	Yes	LOC100826196	U3 small nucleolar ribonucleoprotein protein IMP3-like
	11_10383	AY136627	147.37	Yes	Ha1	plasma membrane P-type proton pump ATPase
	12_30942	GQ169685	147.37	Yes	GS2	plastid glutamine synthetase 2
	11_20422	XM_003560743	28	Yes	LOC100820964	microsomal glutathione S-transferase 3-like

4H						isoform 1
	SCRI_RS_157832	XM_003560569	30	Yes	LOC100840876	vam6/Vps39-like protein-like
6H	SCRI_RS_143317	XM_003570599	72.9	No	LOC100831957	RINT1-like protein-like
	SCRI_RS_206976	XM_003570630	74.6	No	LOC100841641	microtubule-associated protein TORTIFOLIA1-like

CHAPTER 4

TWO GENOMIC REGIONS CONTRIBUTE DISPROPORTIONATELY TO POPULATION STRUCTURE IN WILD BARLEY

Genetic differentiation in natural populations is driven both by geographic distance and by ecological or physical features within and between natural habitats that reduce migration. The primary population structure in wild barley differentiates populations east and west of the Zagros Mountains. The genetic differentiation is greatest in two genomic regions, on linkage groups 2H and 5H. Genetic markers in these two regions demonstrate the largest difference in frequency between the primary populations and have the highest informativeness for assignment to each population. Previous cytological and genetic studies suggest there are chromosomal structural rearrangements (inversions or translocations) in these genomic regions. Environmental association analyses identified an association with both temperature and precipitation variables on 2H and with precipitation variables on 5H.

4.1 Introduction

The wild progenitors of major crops have long been recognized as a valuable genetic resource (cf. Harris, 1990). Natural populations of crop wild relatives have the potential to serve as a source of alleles that contribute to favorable agronomic traits, including cold or drought tolerance (Volis *et al.*, 2002a; Volis *et al.*, 2004) and improved disease resistance (Fetch *et al.*, 2003). Plant germplasm repositories have made substantial investments in the preservation of accessions of both crops and their wild relatives (Schoen and Brown, 2001). Modern genetic approaches have increased the value of these resources as quantitative trait locus (QTL) mapping, association studies, and molecular population genetic studies have combined to uncover specific alleles associated with traits of potential value for crop improvement (Ross-Ibarra *et al.*, 2007; Takeda and Matsuoka, 2008).

Wild barley presents an especially valuable source of potentially useful genes because of its broad geographic distribution and ecological adaptation, spanning ~3500 km east to west from the Levant (the eastern Mediterranean) and Anatolia (present day Turkey) to Central Asia, occurring across much of southwestern Asia. Volis *et al.* (2002b) identified four wild barley ecotypes, which exhibit difference in their level and patterns of phenotypic plasticity. In common garden studies, water stress caused a greater plastic response in the desert ecotype than in the Mediterranean ecotype, whereas for nutrient stress, plasticity was higher in the Mediterranean ecotype than in the desert ecotype (Volis *et al.*, 2002c). The observed differences in ecotype plasticity or local

variation in phenotypic traits is due to environmentally induced local adaptation (Volis *et al.*, 2000; Volis *et al.*, 2002b). Moreover, alleles from wild barley (*Hordeum vulgare* ssp. *spontaneum*) have been used in barley breeding programs for cultivated barley improvement. QTL analyses in an advanced backcross double haploid population derived from a cross between a barley cultivar and a wild barley accession show that wild barley harbors valuable alleles which can improve yield (von Korff *et al.*, 2006; Schmalenbach *et al.*, 2009), malting quality traits (von Korff *et al.*, 2008) and strongly reduce disease symptoms (von Korff *et al.*, 2005). QTL analysis of a recombinant inbred line population and an advanced backcross population derived from crosses between a barley cultivar and a wild barley accession reveals that the wild barley accession contains resistant alleles to multiple fungal pathogens (Yun *et al.*, 2005; Yun *et al.*, 2006). Recently, association and candidate gene resequencing studies have begun to uncover evidence that alleles contributing to agronomically important phenotypes, such as flowering time, may have been introduced from geographic regions outside the initial region of barley domestication, with locally adaptive variants contributing to successful cultivation at higher latitudes (Jones *et al.*, 2008).

Examination of population structure in crop wild progenitors can result in better understanding of the number and geographic region of domestication events (cf. Morrell and Clegg, 2007), which is fundamental to understanding the processes that drove human domestication of plants. The potential to fully exploit genetic variation in the wild relatives of a crop depends, in part, on determining which portions of the range of wild

progenitors have and have not contributed to diversity in the domesticates (Zohary and Hopf, 2000).

Previous studies of wild barley genetic diversity have limitations, such as restricted sampling range, the use of less informative protein data, or limited DNA-based data (Nevo *et al.*, 1986; Morrell and Clegg, 2007). Analysis of 27 isozyme loci in 2,125 individuals sampled from Israel, Turkey and Iran suggested geographic differentiation in wild barley populations (Nevo *et al.*, 1986). Genetic assignment analysis based on resequencing of 18 loci in a sample of 25 to 45 wild barley individuals identified a primary geographic partition of samples into regions east and west of the Zagros Mountains (Morrell and Clegg, 2007). The assignment analysis suggested that additional, geographically distinct wild barley subpopulations might be identified if a larger number of samples were considered (Morrell and Clegg, 2007; Saisho and Purugganan, 2007).

The natural range of wild barley includes a variety of environmental conditions, from arid regions in Central Asia and the Syrian desert to relatively high rainfall coastal regions along the Mediterranean; and from the cold environments in the Zagros Mountains, the western reaches of the Himalayas, and the Iranian plateau to relatively warm lowland Mediterranean coastal regions. By associating the geographic distribution of sequence polymorphisms in georeferenced samples with environmental variables, genomic regions and individual polymorphisms correlated with differences in drought or cold tolerance or other environmental factors may be identified for further investigation.

The questions we seek to explore in this paper are what ecological factors are most associated with observed population structure, how does population structure affect allele

frequency differentiation genome-wide, and how can we use this information to improve conservation and utilization of wild barley genetic diversity? We report the examination of geographic structure and genetic differentiation using 3,072 SNPs (Close *et al.*, 2009) genotyped in a sample of 318 wild barley accessions. We detect two primary populations of wild barley separated by the Zagros Mountains and three subpopulations within each of the two primary populations. Comparison of the two primary populations reveals two pericentromeric regions on the long arms of 2H and 5H that are associated with much of the geographic differentiation in allele frequencies observed in wild barley. The genetic variation in these genomic regions suggests cryptic chromosomal structural rearrangements. Environmental association analyses reveal strong association between these genomic regions and precipitation or temperature.

4.2 Materials and Methods

4.2.1 Materials

The 318 sampled wild barley accessions are known as the Wild Barley Diversity Collection (WBDC) (Steffenson *et al.*, 2007). WBDC accessions were selected to be representative of the geographic range of wild barley, accounting for multiple ecogeographic features (e.g., latitude, elevation, temperature range, and rainfall). The majority of accessions (77.4%) are from the Fertile Crescent, with the balance from Central Asia (15.7%), North Africa (3.8%), and the Caucasus region (2.8%). Individual accessions were self-fertilized for three generations to create inbred lines.

4.2.2 Genotypic data

The WBDC accessions were genotyped using the Illumina Golden Gate Genotyping Assay with two Barley Oligo Pool Assay (BOPA) chips (BOPA1 and BOPA2), each including 1,536 SNPs (Close *et al.*, 2009). The SNPs were discovered by comparison of DNA sequence from expressed sequence tags and sequenced PCR amplicons, derived principally from one wild barley accession and eight malting barley cultivars, primarily from Europe and the United States (Close *et al.*, 2009).

The program ALCHEMY (Wright *et al.*, 2010) was used to generate machine-scored automated genotype calls. The program incorporates estimated inbreeding coefficients for each sample to improvement accuracy of genotype estimation. Unlike programs such as

GenomeStudio, ALCHEMY does not assume Hardy-Weinberg Equilibrium (HWE) genotypic frequencies at each SNP. Inbreeding results in genotypes that depart from HWE expectations. ALCHEMY is based on a Bayesian model of the raw intensity data and can accurately call genotypes. We used three approaches to verify the accuracy of genotype calls. First, WBDC355 (OUH602) has been genotyped separately, with variants segregating in a mapping population (OUH602 by the cultivar Harrington) (Sato *et al.*, 2009b; Muñoz-Amatriaín *et al.*, 2011). Second, genotypes from two lines (WBDC218 and WBDC228) were estimated from RNA-Seq (see below for details of RNA-Seq data processing), so these data were used for validation. Third, all SNPs on BOPA1 were also called manually in GenomeStudio. Only 5% of SNPs have posterior probability of genotype calls < 0.95 from ALCHEMY. We considered these SNPs as missing data.

Before using ALCHEMY for SNP calling, SNPs in BOPA1 and BOPA2 with strong compression or multiple clusters were removed. Subsequent to initial SNP calling, the following quality control steps were applied to the genotyping data. First, we eliminated SNPs that were monomorphic in the wild barley sample. Second, we removed all SNPs that included ≥15% missing data, based on the rationale that large amounts of missing data at a SNP could be associated with inaccurate genotypes. Finally, observed heterozygosity was used as an additional quality control measure. Wild barley is a highly self fertilizing species (Brown *et al.*, 1978) and the WBDC accessions have been subject to three rounds of inbreeding. SNPs with observed heterozygosity >10% were removed on the rationale that the genotypes were likely in error. SNPs with centromeric genetic map positions were identified based on the consensus genetic map of Muñoz-Amatriaín

et al. (2011). In the inbred WBDC lines, observed heterozygosity was extremely low (0.2%), thus we treat the data as haploid genotypes.

SNPs were annotated using the program SNPMeta (Kono *et al.*, 2013). Annotation for each SNP included GenBank ID, gene short name, whether the SNP occurs in coding or noncoding sequence, the SNP position within a codon, and determination if the SNP is silent or induces an amino acid replacement. Among the 3,072 BOPA SNPs, 2,508 were annotated with 338 derived from named genes.

In addition to SNP data, we examined 29 microsatellite loci in all WBDC accessions (Ramsay *et al.*, 2000; Li *et al.*, 2003; Varshney *et al.*, 2007). Microsatellites offer the advantage of reduced ascertainment bias, because the microsatellites are selected to be polymorphic, but the selection process is not conditional on the presence of an individual polymorphism (Haasl and Payseur, 2010). Microsatellite locus names, repeat number, repeat unit length, total size, and heterozygosity for each microsatellite can be found in Table S4.1. This information was obtained from GrainGenes: A Database for Triticeae and Avena (Matthews *et al.*, 2003). The program R_{ST} Calc (Goodman, 1997) was used to compare allele frequency differences among partitions of the sample, assuming a stepwise mutation model (Slatkin, 1995).

4.2.3 RNA-Seq data processing

RNA-Seq reads from *H. bulbosum* accession Cb2920/4 (used as outgroup to infer ancestral state), and wild barley WBDC218 and WBDC228 were trimmed for adapter contamination with Scythe (https://github.com/vsbuffalo/scythe) and then aligned to the

Morex draft sequence (Mayer *et al.*, 2012) with Bowtie 2 (Langmead and Salzberg, 2012). We adjusted read mapping parameters to accommodate the expected divergence between *H. vulgare* lines (~1%) and between *H. vulgare* and *H. bulbosum* (~3%) (Morrell *et al.*, in press). Alignments were processed with Samtools (Li *et al.*, 2009) and the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010; DePristo *et al.*, 2011) according to the GATK best practices (http://www.broadinstitute.org/gatk/guide/best-practices). For realignment around indels, we used a set of high-confidence indels reported in Sanger resequencing datasets (Caldwell *et al.*, 2006; Morrell *et al.*, 2006; Morrell and Clegg, 2007). We then extracted the base calls at each BOPA SNP location using tools from the GATK.

4.2.4 Geographic differentiation

We examined geographic structure within the range of wild barley with Bayesian genetic assignment implemented in the program STRUCTURE (Pritchard *et al.*, 2000; Falush *et al.*, 2003). STRUCTURE assumes there are *K* clusters for the samples and assigns the individuals to clusters based on distinct allele frequencies. We treat individual samples as haploid and explored both admixture and no admixture models and models with correlated and uncorrelated allele frequencies with $K = 2 - 10$ clusters. Because a model with no admixture and uncorrelated allele frequencies resulted in higher likelihoods, it was used for the final analyses. For each value of *K*, we used 10 replicate runs, with a burn-in length of 100,000 iterations and a run length of 100,000 iterations.

Haplotypes were defined based on five adjacent SNPs and used both for comparison of haplotype diversity and as input for genetic assignment analysis. Combining markers into haplotypes results in multi-allelic data that can improve inference of population structure (Haasl and Payseur, 2010; Gattepaille and Jakobsson, 2012). To infer missing data, we used the program fastPHASE (Scheet and Stephens, 2006) with 20 random starts and 25 iterations of the Expectation-Maximization algorithm. The number of SNPs within each haplotype was determined based on optimal numbers from the simulation study of Gattepaille and Jakobsson (2012) and the total number of SNPs in our samples. To deal with label switching (where cluster names change between replicate runs) and with true multimodality (where individual samples switch clusters in replicate runs), we used CLUMPP (Jakobsson and Rosenberg, 2007) to summarize assignment results across replicate runs. The program Infocalc was used to calculate the informativeness for assignment (*In*) (Rosenberg *et al.*, 2003) for each haplotype using the clusters identified by STRUCTURE. Informativeness for assignment identifies the information content for genetic assignment for markers based on the degree to which each locus (or haplotype segment) contributes to distinction among populations (Rosenberg *et al.*, 2003).

The prcomp function in R (R Development Core Team, 2011) was used to perform principal component analysis (PCA). Each WBDC accession was assigned to clusters based on significant principal components (PCs) using the Ward clustering method in the R hclust function. To compare the similarity between genetic variation on PCA plot and geographic maps of sample locations, we used Procrustes analysis to find a rotation that maximizes the similarity (Wang *et al.*, 2012).

Within and among inferred clusters, we estimated hierarchical F-statistics (Yang, 1998) using the Hierfstat package (Goudet, 2005) in R. We calculated summary statistics, including number of segregating sites, and number of private alleles in each cluster using tools from the libsequence library (Thornton, 2003), and estimated average pairwise SNP diversity within each cluster in the R package ape (Paradis *et al.*, 2004). Rarefaction was used to analyze allelic diversity across populations while correcting for sample size differences using the program ADZE (Szpiech *et al.*, 2008).

Linkage disequilibrium (LD) as measured by r^2 (Hill and Robertson, 1968) was calculated for all possible pairwise comparisons on each linkage group based on SNPs with minor allele frequency (MAF) >5%. The LDheatmap package (Shin *et al.*, 2006) was used to generate plots of LD relative to genetic distance.

4.2.5 Local adaptation

To identify genomic segments potentially contributing to local adaptation, we divided samples into the two primary clusters (the Eastern and Western populations) and six clusters identified by STRUCTURE and calculated F_{ST} (Weir and Cockerham, 1984). The assumption is that SNPs linked to genomic regions contributing to local adaptation will show greater allele frequency divergence (higher F_{ST}) than those affected only by demography (Cavalli-Sforza, 1966; Lewontin and Krakauer, 1973).

Environmental variables, including altitude, monthly precipitation, monthly maximum and minimum temperature, and 19 additional bioclimatic variables were downloaded from www.worldclim.org (Hijmans *et al.*, 2005). DIVA-GIS (Hijmans *et al.*,

2001) was used to extract climate data at 5 arc-min (~10 km) resolution for each sample. We focused on measurements within the growing season for wild barley, from late autumn to spring, so we removed environmental variables related to summer temperature and precipitation. The environmental variables used are listed in Table S4.2. Environmental variables were scaled to a mean of 0 and standard deviation of 1 and grouped into principal components. The significant PCs were used as environmental variables for Bayenv (Coop *et al.*, 2010).

We used Bayenv to identify the correlation between SNPs and environmental variables. Bayenv requires population information to account for population structure. Bayenv makes use of allele frequency within a set of samples representing a localized environment to correct for population structure in environmental association. We used PCA to group all samples into 17 clusters with a mean sample size of 17 accessions. The number of optimal stratifications in our data was determined using Velicer's minimum average partial test (Shriner, 2011). We used all SNPs to construct the covariance matrix. Two independent runs of 30,000 iterations were compared to control for convergence and the final covariance matrix is the mean of these two independent runs. Then Bayenv was used to estimate the Bayes factor for each SNP with each environmental variable using 50,000 iterations. SNPs were considered candidates contributing to local adaptation if they have an average Bayes factor above the 95th percentile genome-wide for five separate runs.

Because the wild barley samples cover a large continuous geographic range, spatial ancestry analysis (SPA) (Yang *et al.*, 2012) was used to model allele frequency change

for each SNP as a function of the location of the individual in geographic space. SNPs contributing to local adaptation potentially have larger gradients in allele frequency, reflected in high SPA score (Yang *et al.*, 2012). SNPs were considered candidates for local adaptation if their SPA score is above the 95th percentile genome-wide.

SNPs that are outliers (above the 95th percentile genome-wide) in the F_{ST}, environmental association, or SPA analyses can be considered candidate variants either causative, or more likely linked to, loci involved in local adaptation. Enrichment analysis of candidates among genic versus non-genic and non-synonymous versus synonymous was performed by resampling the number of SNPs in each candidate list randomly from the genome 1,000 times. Enrichment analyses included sets of SNPs genome-wide, centromeric regions and the two regions that are putative structural rearrangements (inversions or translocations). This generated a distribution of ratios of the number of genic to the number of non-genic SNPs and ratios of the number of non-synonymous to the number of synonymous SNPs in each analysis, which can be compared to the observed ratios from the data.

4.3 Results

Initial screening of relatedness among accessions identified 30 wild barley accessions with either a large genetic distance from the majority of accessions or that appear to be duplicated within the sample. The large genetic distance was also associated with a high degree of identity to barley landraces genotyped on the same platform. The increase in genetic distance appears to result from ascertainment bias due to the discovery of barley SNPs primarily among cultivated samples, which results in a larger number of segregating polymorphisms in barley landrace accessions than in wild barley (Russell *et al.*, 2011). The 30 accessions were excluded from further analysis because they constitute duplicate accessions or because genotypic composition suggests they could be either feral barley accessions or were subject to recent introgression. Four accessions do not have known latitude and longitude of origin, so they were removed from analyses, thus 284 wild barley accessions from the WBDC were used in this study. After all quality control measures, 2,330 SNPs were assayed in each accession.

Microsatellite loci were extremely polymorphic among wild barley accessions with an average of 20 alleles per locus. Allele size for individual microsatellites is normally distributed and thus accords with a stepwise mutation model (Valdes *et al.*, 1993).

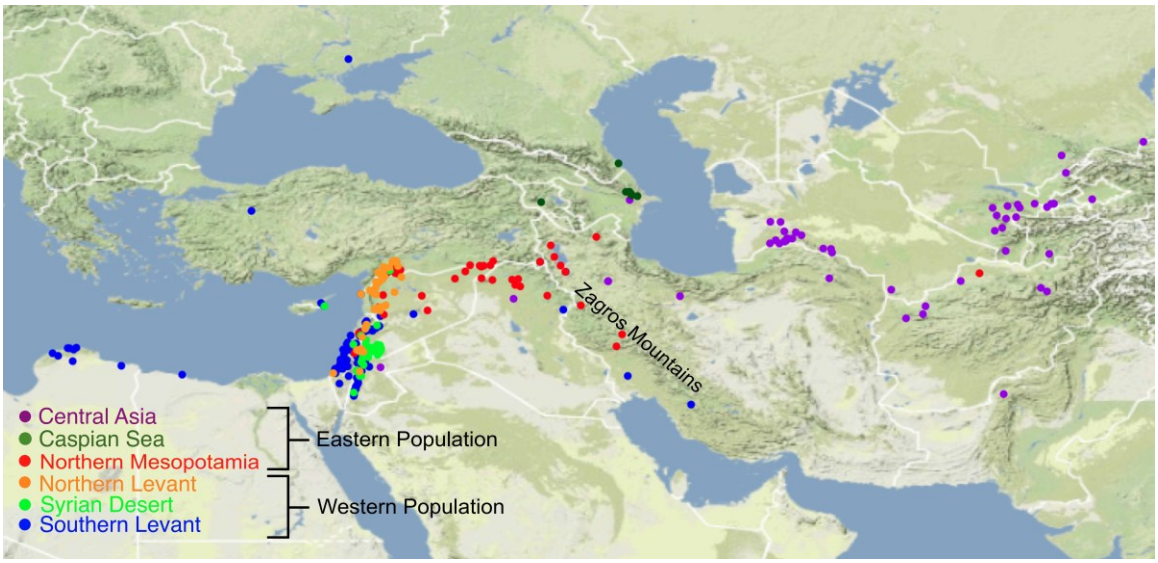
4.3.1 Population structure

A Mantel test identifies a positive correlation between geographic distance and genetic distance (Mantel statistic: 0.37, significance at 0.001), consistent with isolation by distance among these wild barley accessions.

STRUCTURE results based on single SNP and five-SNP haplotypes for $K = 2$ identified populations east and west of the Zagros Mountains. Samples from a broad portion of the range, including Central Asia (particularly east of the Caspian Sea), the Iranian Plateau and most samples in Northern Mesopotamia (modern Northern Iraq and Syria) form an Eastern population, and samples west of the Zagros Mountains, including samples from the Levant, and around the Mediterranean to North Africa form a Western population (Figure 4.1A). For genetic assignment based on individual SNPs, there are six accessions from Central Asia assigned to the Western population (data not shown). Thus genetic assignment from haplotype data show greater consistency with geographic location of origin. The broad scale geographic patterns identified from SNP and 5-SNP haplotypes are not readily reflected in genetic assignment based on the 29 microsatellite loci. The individual alleles are relative rare and have a low informativeness for assignment, an issue attributable to the high levels of polymorphism in these microsatellites (Table S4.1).

For STRUCTURE analysis based on haplotype data, when $K = 3$, a group of accessions from the Syrian desert region becomes an independent cluster from the Western population. As K increases to 4, the samples along the east coast of Mediterranean split into two groups, one in the north (Northern Levant), and the other in

(A)


Central Asia
Caspian Sea
Northern Mesopotamia
Northern Levant
Syrian Desert
Southern Levant
Eastern Population
Western Population
Zagros Mountains

(B)

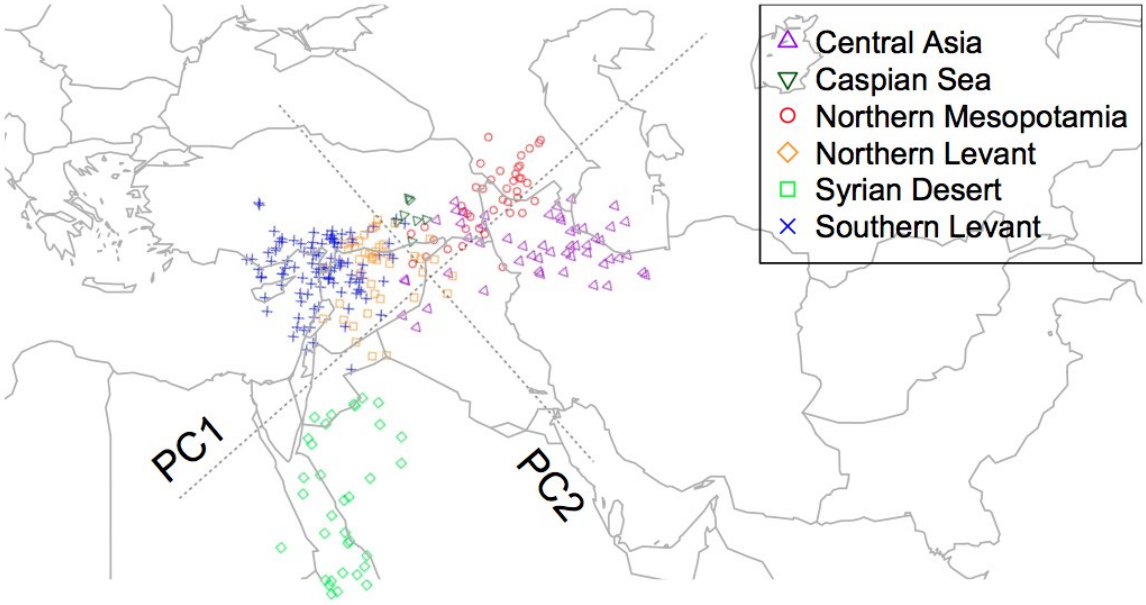

Central Asia
Caspian Sea
Northern Mesopotamia
Northern Levant
Syrian Desert
Southern Levant
PC1
PC2

Figure 4.1 (A) Population structure in wild barley. Each of the six colors represents one of the six subpopulations. Three different subpopulations are nested in the Eastern and Western populations respectively. (B) Procrustes-transformed PCA plot of genetic variation in wild barley.

the south (Southern Levant). The Eastern population begins to differentiate as *K* increased to 5. The samples in Central Asia are separated from samples from Northern Mesopotamia and those from west of the Caspian Sea. With *K* = 6, the seven samples from the Caspian Sea become an independent cluster (Figure 4.1A). For the present sample, *K* = 6 provides a clear distinction among populations and the genetic assignment is not constrained by very small sample size within individual clusters. The average informativeness for assignment for individual SNPs is 0.03 for *K* = 2 and 0.10 for *K* = 6; 0.14 for *K* = 2 and 0.47 for *K* = 6 based on haplotypes (Figure S4.1).

Among the six clusters, three fall within the Eastern population (Central Asia, Caspian Sea, and Northern Mesopotamia) and three into the Western population (Northern Levant, Syrian Desert, and Southern Levant), constituting hierarchical population structure nested within the major Eastern and Western populations. There is both a strong effect of individuals within subpopulations (p-value = 0.01, 100 permutations) and of subpopulations within populations (p-value = 0.04, 100 permutations). Population structure is best explained by a six subpopulation model (variance components = 17.7%), in contrast the two population model provides a poorer fit to the data (variance components = 10.6%). F_{ST} values are also higher at the subpopulation level than the population level (Table 4.1). Within the two higher-level populations, 20.5% of the variance can be explained by the three subpopulations within

the Eastern population while 13.2% can be explained by the three subpopulations within the Western population.

Table 4.1 Hierarchical F-statistics comparing different levels of the hierarchical population structure.

	Population	Subpopulation
Total	0.042	0.191
Population	0.000	0.155

In the PCA examination of population structure, the first PC separates all samples into the Eastern and Western populations. When adding the second PC, all the six subpopulations are clearly differentiated. A PCA plot of genetic variation closely reflects geography in the population after a 42.44º counterclockwise rotation of the PCA plot using Procrustes analysis. The boundary of the Eastern and Western population roughly parallels the Zagros Mountains (Figure 4.1B). The Syrian Desert population is noteworthy in showing greater PC distance within and among samples, possibly owing to greater genetic drift within the population. A focal F_{ST} analysis indicates a higher mean F_{ST} value for this population (0.09) than for the other two Western populations (0.05 and 0.04).

4.3.2 Population comparison

The average genome-wide F_{ST} between the Eastern and Western populations is 0.07. The boundary of these two populations is close to the Zagros Mountains (Figure 4.1A), which forms a potential barrier to migration (Lin *et al.*, 2001; Morrell *et al.*, 2003). The

most northerly portion of the Zagros Range is at ~48º E longitude, trending from the northwest to the southeast, so we also compared the allele frequency differentiation

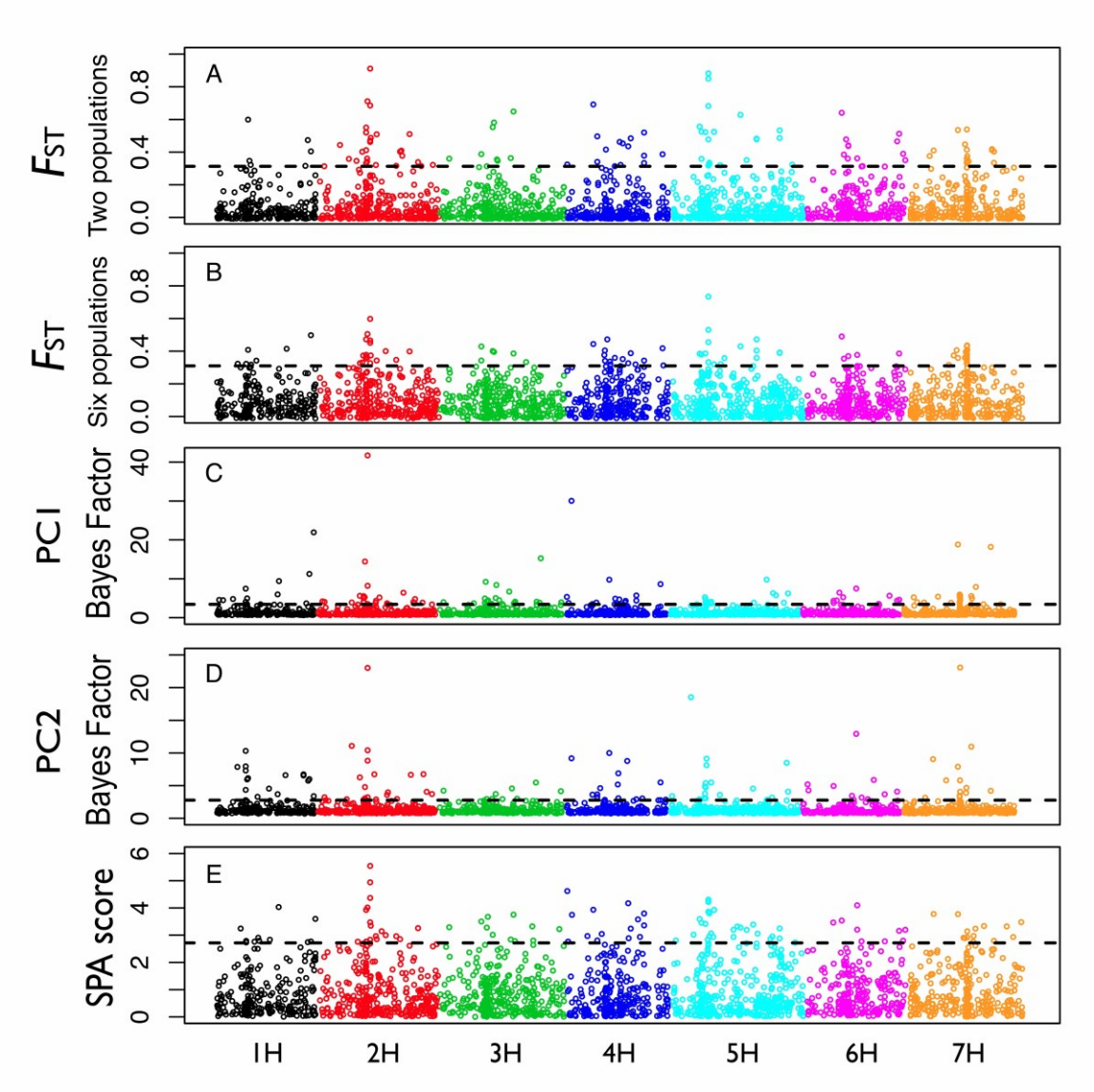


Figure 4.2 (A) *F*ST between the Eastern and Western populations. (B) Pairwise *F*ST based on all six subpopulations. (C) Bayes factors for correlation between allele frequencies and PC1. (D) Bayes factors for correlation between allele frequencies and PC2. (E) SPA score genome-wide from spatial analysis.
In (A) and (B), the dashed dotted line is the 95th percentile of *F*ST genome-wide. In (C), (D), and (E), the 95th percentile of the distribution of Bayes factors or SPA scores is indicated by a horizontal dashed line.

between the two populations east and west of 48º E. For this geographic contrast, the average F_{ST} genome-wide is 0.06 and the correlation (r^2) between this partition based on the Zagros Mountains and the previous partition based on genetic assignment analysis is 0.473, which supports the hypothesis that the Zagros Mountains act as a natural barrier that bisects wild barley into the Eastern and Western populations. For the 29 microsatellite loci, the average R_{ST} between the Eastern and Western populations is 0.15.

SNPs with high F_{ST} values between the Eastern and Western populations (Figure 4.2A) or among all the six subpopulations (Figure 4.2B) are enriched in two genomic regions, one on linkage group 2H, from ~67 cM to 74 cM, the other on 5H, from ~47 cM to 52 cM. For the 2H and 5H regions, the mean F_{ST} is 0.20 (56 SNPs) and 0.17 (32 SNPs) respectively versus a genome-wide F_{ST} = 0.07. An F_{ST} versus minor allele frequency plot is shown in Figure S4.2. F_{ST} outliers have very high minor allele frequencies.

Many SNPs or haplotypes genome-wide with high informativeness for assignment fall in these two high F_{ST} regions (Figure S4.1). Among all SNPs above 95th and 99th percentile (I_n = 0.31 and 0.42), 33% and 52% are in these two regions (Figure S4.1). However, when we perform genetic assignment analysis after masking these two high F_{ST} regions, the probability of assignment for each wild barley accession into the Eastern and Western population is nearly identical for all but one accession. This result reflects the relatively high informativeness for assignment observed for SNPs within pericentromeric regions on all linkage groups (Figure S4.1). Therefore, population structure in these two high F_{ST} regions is similar to the genome-wide pattern and the

genome-wide pattern is driven by a high degree of differentiation in pericentromeric regions.

The joint unfolded site frequency spectrum demonstrates that there are more rare variants in the Western population than Eastern population (Figure S4.3). Percent pairwise differences are lower in the Eastern population than in the Western population (Table 4.2). There are more private SNPs in the Western population than in the Eastern population (430 versus 86) (Table 4.2). Because the sample size is different between the Eastern and Western populations, we used rarefaction to correct for sample size. Despite the correction, both the mean number of distinct alleles per locus (Figure S4.4A) and the mean number of private alleles per locus (Figure S4.4B) are higher in the Western than the Eastern population. The Southern Levant subpopulation has the highest values for the

Table 4.2 Diversity summary statistics for the two populations and six subpopulations
The summary statistics include the sample size, number of segregating sites, number of private alleles, percent pairwise difference with standard deviation (SD) and microsatellite expected heterozygosity.

Population	Size	# segregating sites	# private alleles	Percent pairwise difference (SD)	Microsatellite expected heterozygosity
Eastern	101	2196	86	0.20 (0.04)	0.740
Caspian Sea	7	1146	2	0.10 (0.02)	0.672
Central Asia	53	2027	22	0.19 (0.04)	0.734
Northern Mesopotamia	41	1975	13	0.17 (0.03)	0.717
Western	183	2285	430	0.23 (0.03)	0.742
Northern Levant	42	2033	19	0.21 (0.02)	0.740
Southern Levant	107	2197	49	0.22 (0.02)	0.736
Syrian Desert	34	1916	6	0.16 (0.03)	0.722

number of segregating sites, and number of private SNPs while the Caspian Sea subpopulation has the lowest values for these summary statistics (Table 4.2).

4.3.3 Structural rearrangements

The high F_{ST} regions on 2H and 5H are potentially attributable to chromosomal structural variants. Population genetic variation, particularly patterns of LD, can be strongly suggestive of structural variation (Huynh *et al.*, 2011; Long *et al.*, 2013). The average pairwise LD in the high F_{ST} region on 2H ($r^2 = 0.063$) is higher than other regions of 2H ($r^2 = 0.012$). The average pairwise LD in the high F_{ST} region on 5H ($r^2 = 0.068$) is also higher than other regions on 5H ($r^2 = 0.014$). The 5-SNP segment with the lowest haplotype number (4) on 2H is within the high F_{ST} region (Figure S4.5). As on other linkage groups, the segment with the lowest haplotype number on 2H and 5H is within the centromeric region (Figure S4.5). The centromeric region on 2H overlaps with the high F_{ST} region while it is distal to the high F_{ST} region on 5H (Figure S4.5). The observed LD, F_{ST} and haplotype number patterns are consistent with recent positive selection and/or chromosome structural rearrangements, but the average pairwise r^2 is lower than that observed in a chromosomal inversion in the wild ancestor of maize, where the average pairwise r^2 is 0.24 in an ~50 Mb region (Fang *et al.*, 2012).

The high F_{ST} region on 2H occurs in the same approximate chromosomal location as a chromosomal rearrangement identified in an eastern wild barley accession based on meiotic pairing studies (Konishi and Linde-Laursen, 1988). Konishi and Linde-Laursen (1988) report a reciprocal translocation with 4H in a sample from Turkmenistan with the

breakpoints of the translocation near the centromere on 2H. Using a three point linkage test, Ramage and Suneson (1961) identify evidence of both an inversion and translocations on the long arm of 5H (Ramage and Suneson, 1961).

4.3.4 Evidence for local adaptation

SNPs that occur as outliers in the F_{ST} analysis may indicate genomic regions involved in local adaptation. Annotation information for SNPs that are above the 95th percentile of F_{ST} between the Eastern and Western populations is listed in Table S4.3.

Environmental association analysis was used to identify genetic polymorphisms potentially involved in local adaptation. PCA reveals two major clusters of environmental variables (Figure S4.6). The first two PCs explain 80% of the total variance (Figure S4.7). The first PC includes most temperature variables while most precipitation variables and altitude are in the second PC (Table S4.4).

Environmental association analysis reveals that SNPs associated with both PC1 and PC2 (above the 95th percentile) are distributed on all linkage groups (Figure 4.2C, D). The high F_{ST} region on 2H is highly associated with both PC1 and PC2 (Figure 4.2C, D). The high F_{ST} region on 5H is also associated with PC2 (Figure 4.2D). The annotation information for SNPs that are above the 95th percentile of association with PC1 and PC2 is listed in Table S4.5.

SPA analysis reveals that ~20% of the SNPs in the high F_{ST} regions on both 2H (11 out of 56) and 5H (7 out of 32) show strong geographic gradients in allele frequencies as their SPA scores are above the 95th percentile (2.72) (Figure 4.2E). The SNP with the

highest SPA score (5.54) falls in the high F_{ST} region on 2H. There are 24 SNPs with SPA score above the 99th percentile (3.75) and nearly half of these SNPs (11) are in these two high F_{ST} regions on 2H and 5H. The annotation information for the SNPs with SPA scores above the 95th percentile is in Table S4.6.

Enrichment analysis reveals that the candidates in the two putative chromosomal structural rearrangements are enriched for genic SNPs (Figure S4.7A). The candidates in the centromeric regions are enriched for non-synonymous SNPs (Figure S4.7B).

4.4 Discussion

4.4.1 Hierarchical population structure

Wild barley shows strong hierarchical population structure, with primary structure east and west of the Zagros Mountains and three subpopulations identified in both the Eastern and Western populations (Figure 4.1A). Previous studies of sequence diversity in wild barley have identified population structure that strongly differentiates the Eastern and Western wild barley populations (Lin *et al.*, 2001; Morrell and Clegg, 2007; Saisho and Purugganan, 2007). The present study samples a much larger number of accessions and includes only 22 SNPs in common with those sampled in previous studies. Despite the limited overlap of sampled SNPs, genetic assignment with $K = 2$ uncovers a similar pattern. Moreover, the results of this study go beyond previous work by revealing a

division into six subpopulations that explains 7.1% more variance compared with the primary two population division.

A nearly continuous geographic range represents a particular challenge for efforts to identify population structure and geographic discontinuities. However, populations of wild barley are much more common in the western portion of the range and below 1500 m (Zohary and Hopf, 2000), thus the 3000 - 4500 m peaks of the Zagros Mountains and the high elevation regions of the Iranian plateau are disruptions of an otherwise continuous range.

The Western population is more diverse than the Eastern population (Table 4.2) at least in part because the SNP discovery panel was composed primarily of Western cultivars. It should be noted that the discovery panel also included OUH602 (WBDC355), which assigns to the Central Asia (Eastern) wild barley population. Estimates of diversity based on resequencing a more limited set of wild barley samples indicate moderately higher levels of diversity in the Western than in the Eastern wild barley populations (Morrell *et al.*, in press; Morrell and Clegg, 2007). The effect of ascertainment bias on the frequency spectrum depends on the population in which SNPs were discovered (Albrechtsen *et al.*, 2010b). Therefore, we observe more rare alleles in the Western population (Figure S4.3). There is a small but clear effect of ascertainment bias, which leads to an increased estimate of diversity in the Western population, because the Western population is geographically more similar to the discovery panel. The size of a discovery panel is less important than the composition, as long as it is not extremely small (< 4 chromosomes) (Albrechtsen *et al.*, 2010b). Based on a simulation study, a

discovery panel of eight samples with a minimum allele count of three best reflects the design parameters of the BOPA SNPs (Fang *et al.*, 2013).

4.4.2 Two putative chromosome rearrangements

Resequencing studies identified a large degree of heterogeneity in wild barley, in terms of both degree of population structure and levels of nucleotide sequence diversity (Lin *et al.*, 2001; Morrell *et al.*, 2003). Using simple coalescent simulations, Morrell *et al.* (2003) argued that demographic effects alone were insufficient to explain intralocus heterogeneity and that selection, either as selective sweeps at some loci or local adaptation at others was necessary to explain observed patterns of diversity (Wright *et al.*, 2005). Strong genetic differentiation between the Eastern and Western populations for the two regions on 2H and 5H suggest that some of the heterogeneity may result not from selection acting on individual loci, but on structural variants (Figure 4.2). Structural variants are quickly lost due to drift and purifying selection unless they confer a locally adaptive advantage. A structural variant that captures two or more alleles adapted to the local environment has a selective advantage that can cause it to spread (Kirkpatrick and Barton, 2006).

The two genomic regions with high F_{ST}, also have above average levels of LD and low haplotype number (Figure S4.5). Environmental association analysis identifies multiple SNPs in these regions associated with both temperature and precipitation (Figure 4.2). SPA analysis identifies half of the SNPs in these two regions as outliers in terms of dramatic change in allele frequency gradients, as identified by SPA scores above the 99th

percentile. The SPA method is particularly sensitive to SNPs that have steep geographic gradients in allele frequency and is scored based on individual accessions rather than populations (Yang *et al.*, 2012).

The large number of SNPs in these regions identified in multiple analyses indicate these two regions may harbor variants that are locally adaptive, with selection altering the frequency of nearby SNPs through genetic hitchhiking (Nielsen *et al.*, 2005). Given the density of SNPs assayed in the present study and the relatively rapid decay of LD in wild barley (Morrell *et al.*, 2005), the detection of multiple SNPs associated with environmental factors is likely to occur only in regions with suppressed recombination. These two high F_{ST} regions are 5 - 7 cM, potentially including hundreds of genes, thus the patterns observed are likely due to chromosome structural variants, which through the inhibition of genetic exchange between chromosomal rearrangements, have the effect of slowing down migration. Previous cytological and genetic studies also suggest that there are translocations or inversions in these genomic regions (Ramage and Suneson, 1961; Konishi and Linde-Laursen, 1988).

The Eastern accessions occur in a region that on average is more arid than the region occupied by the Western accessions, where most samples were collected from populations along the coastal Mediterranean. The precipitation pattern is reflected in the environmental association analysis. Both of the two putative chromosome rearrangements on 2H and 5H are associated with precipitation variables (Figure 4.2). There are more unique haplotypes in the Western population than that in the Eastern population (Figure 4.3). In the Eastern population, both of the two regions on 2H and 5H are dominated by a

few haplotypes (Figure 4.3), which is potentially consistent with selection favoring these haplotypes.

We also noted that the two putative chromosome rearrangements are not differentiated by large numbers of private SNPs (Figure 4.3) as observed at the largest inversion (*Inv1n*) in teosinte, the wild progenitor of maize (Fang *et al.*, 2012). Moreover, the putative translocation on 2H incorporates the centromere and the putative rearrangement on 5H is close to the centromeric region. Therefore, the patterns observed could be exceptional effects, arising from suppressed recombination in the centromeric regions (Mayer *et al.*, 2012).

4.4.3 Useful wild barley alleles

Identification of functional variation that indicates local adaptation can contribute to sustained crop improvement. The history of crop wild progenitors is orders of magnitude longer than the history subsequent to domestication, so wild populations have been exposed to natural selection for many more generations. Moreover, a domestication bottleneck decreases nucleotide diversity and causes the loss of valuable variants (Eyre-Walker *et al.*, 1998). For these reasons, wild populations are expected to carry many novel nucleotide sequence variants and functional adaptations that are not present in domesticates.

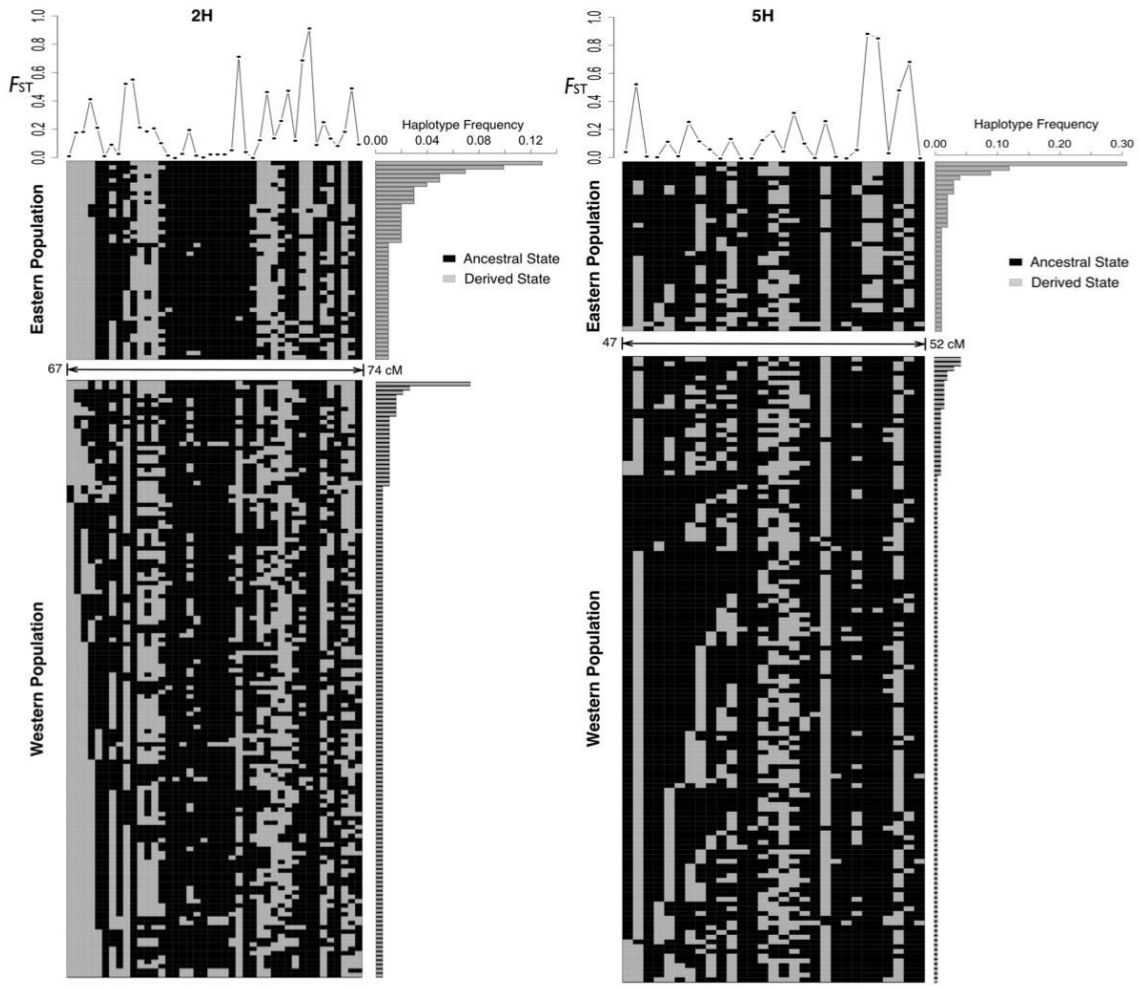


Figure 4.3 Diagram of haplotype diversity in the two putative chromosome structural rearrangements on 2H and 5H.
Haplotypes are divided into the two primary populations identified by STRUCTURE. Each SNP is represented by either the ancestral state (black) or the derived state (gray). The frequency of each of the haplotypes from the Eastern population (top) and the Western population (bottom) is shown on the right. *F*ST of each SNP between these two populations is shown on the top of the haplotype diagram.

We used several approaches to identify nucleotide polymorphisms potentially involved in local adaptation. SNPs that are outliers in the F_{ST}, environmental association, and SPA analyses are potentially linked to loci contributing to local adaptation. The F_{ST} comparison makes the assumption that genetic markers with extreme allele frequency differences among populations may contribute to local adaptation (Lewontin and Krakauer, 1973), but this method has several limitations. First, the Lewontin and Krakauer approach suffers from a number of assumptions, including that all populations diverged at the same time (Nei and Maruyama, 1975). Second, comparison of allele frequency differences with F_{ST} requires the prior identification of populations. Environmental association and SPA analyses complement and are consistent with the F_{ST} results. The results of both analyses support the conclusion that selection has contributed to the observed allele frequency differentiation likely driven by environmental factors or correlated selection pressures (Coop *et al.*, 2010).

A number of SNPs within previously characterized barley loci are outliers in one or more of the analyses reported here. The SNP (12_30850) on *Cbf4* is above the 95th percentile of F_{ST} values for SNP frequencies compared between the Eastern and Western populations. *Cbf4* determines low-temperature tolerance in barley (Francia *et al.*, 2004). From environmental association analysis, one SNP (11_11361) in *Cbf4* and another SNP (11_10989) in a cold-regulated gene *blt14* (Cattivelli and Bartels, 1990; Grossi *et al.*, 1998) are among outliers of PC1, which includes the temperature variables. SPA analysis reveals that SNP 12_30850 in the *Cbf4* locus also shows a strong geographic gradient in allele frequency.

4.4.4 Summary

Much of the population structure observed in wild barley can be accounted for by two large pericentromeric regions on 2H and 5H. These two genomic regions are putative chromosome structural rearrangements that harbor variants that appear to contribute to environmental adaptation. In particular, the SNPs in these chromosomal regions are shown to be associated with temperature and precipitation variables. It will be important to determine how specific genetic variants within these rearrangements are associated with local adaptation. Nevertheless, the identification of genomic regions associated with environmental adaptation suggests an opportunity for crop improvement.

4.5 Supporting Information

4.5.1 Supplementary Figures

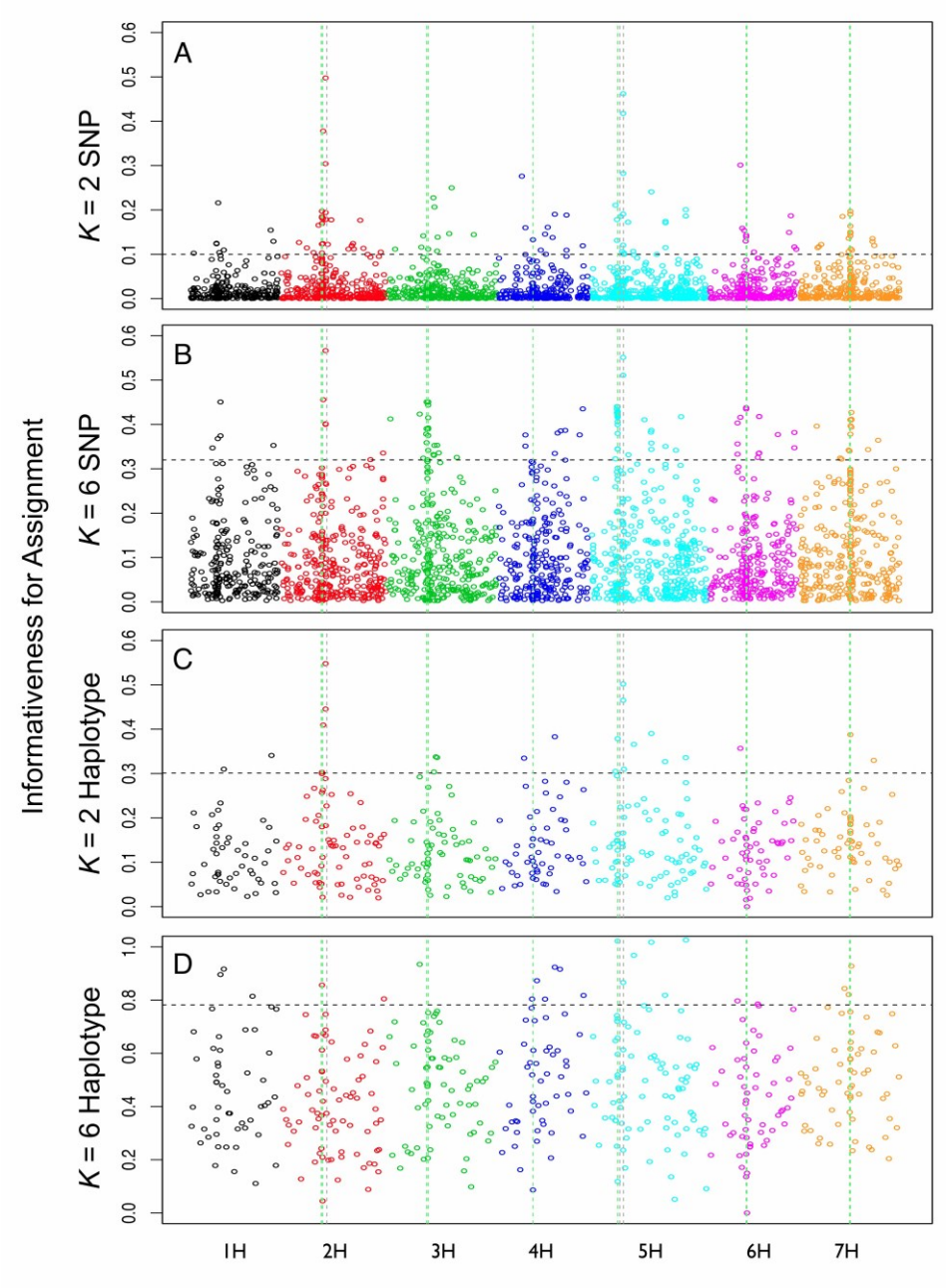


Figure S4.1 The informativeness for assignment for all SNPs (A, B) and 5-SNP haplotypes (C, D) genome-wide based on (A, C) $K = 2$ and (B, D) $K = 6$.
The horizontal dashed line is the 95th percentile. The grey vertical dashed lines delineate the two high F_{ST} regions based on comparison between the Eastern and Western populations. The green vertical dashed lines indicate the centromere regions.

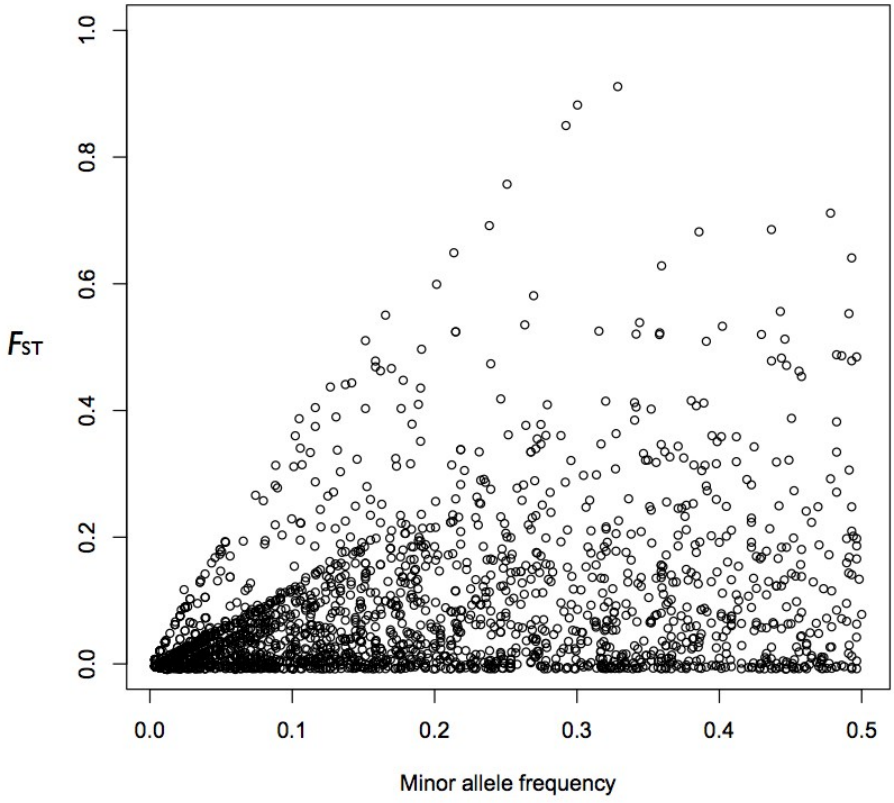


Figure S4.2 F_{ST} genome-wide based on comparison between the Eastern and Western populations versus minor allele frequency from all accessions.

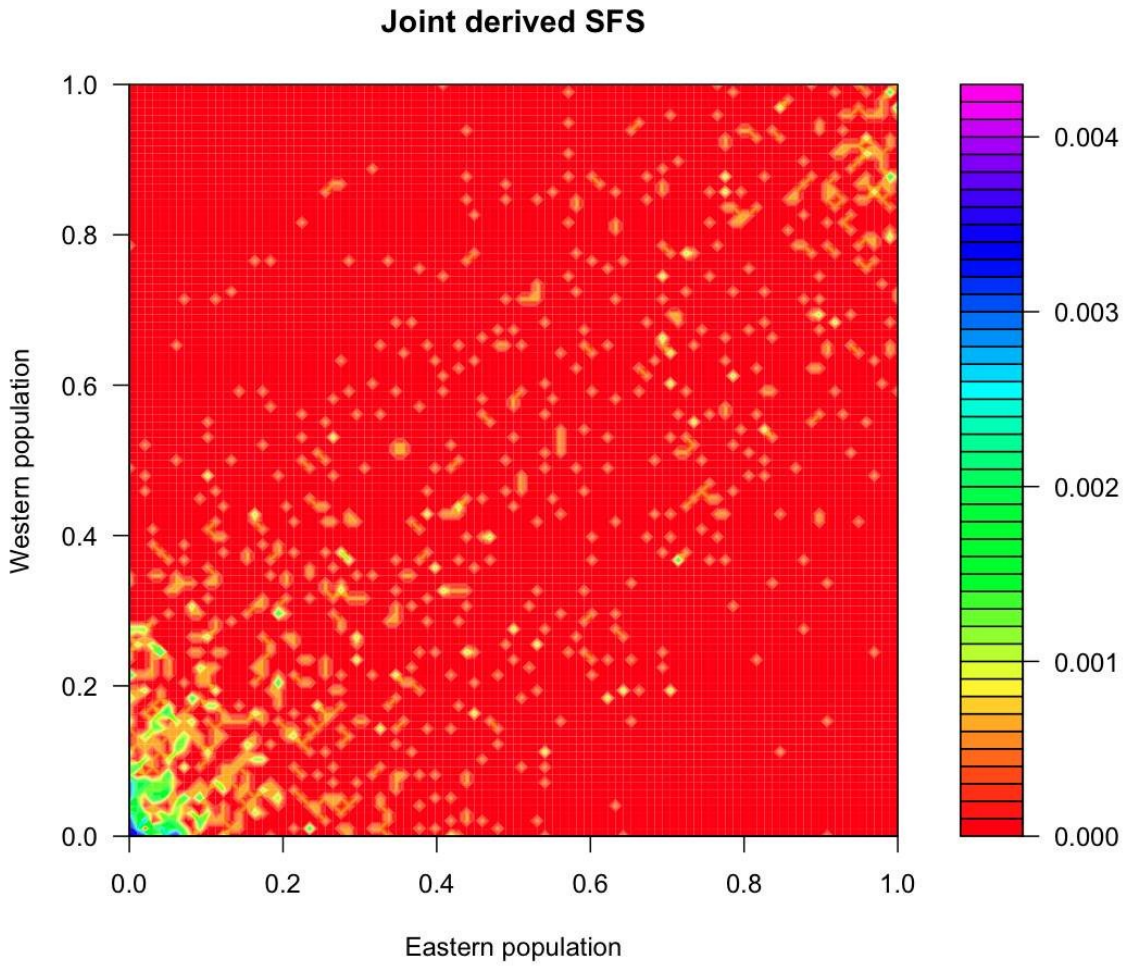
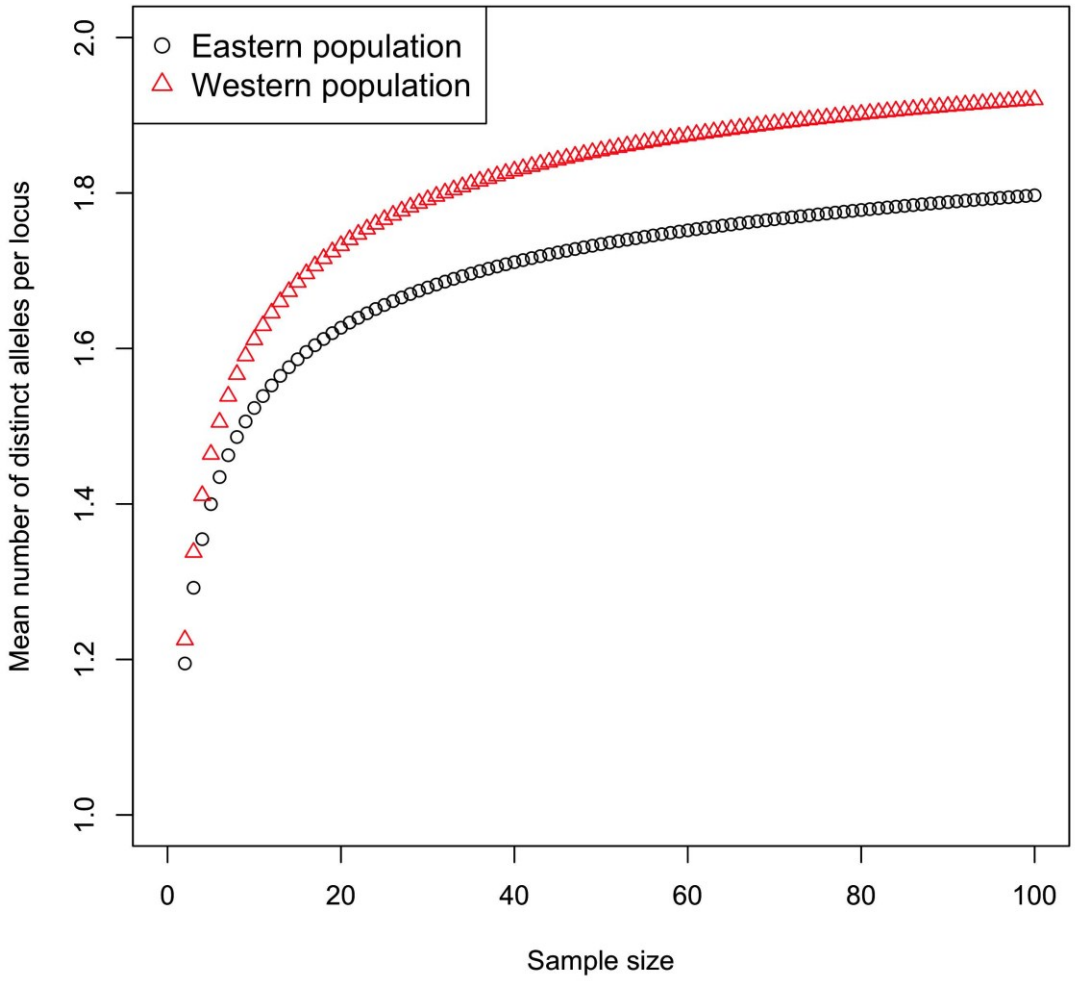


Figure S4.3 The joint unfolded site frequency spectrum based on all accessions from the Eastern population (upper triangle) and Western population (lower triangle).
The comparison includes 1633 SNPs for which the ancestral state could be inferred by comparison to *H. bulbosum* Illumina resequencing data.

(A)


Eastern population
Western population
Mean number of distinct alleles per locus
1.0
1.2
1.4
1.6
1.8
2.0
0
20
40
60
80
100
Sample size

(B)

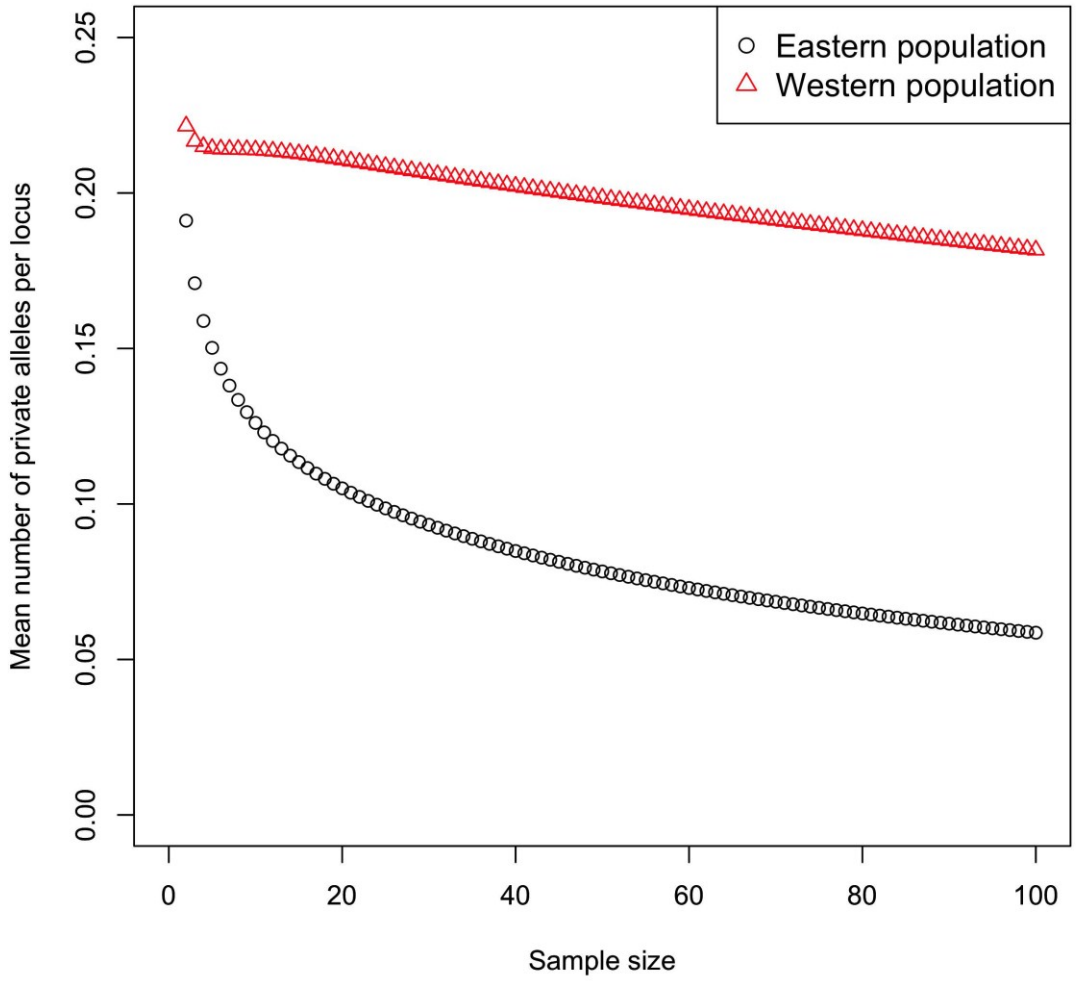


Figure S4.4 Rarefaction analysis comparing nucleotide diversity between the Eastern and Western populations
(A) mean number of distinct alleles per locus and (B) mean number of private alleles per locus.

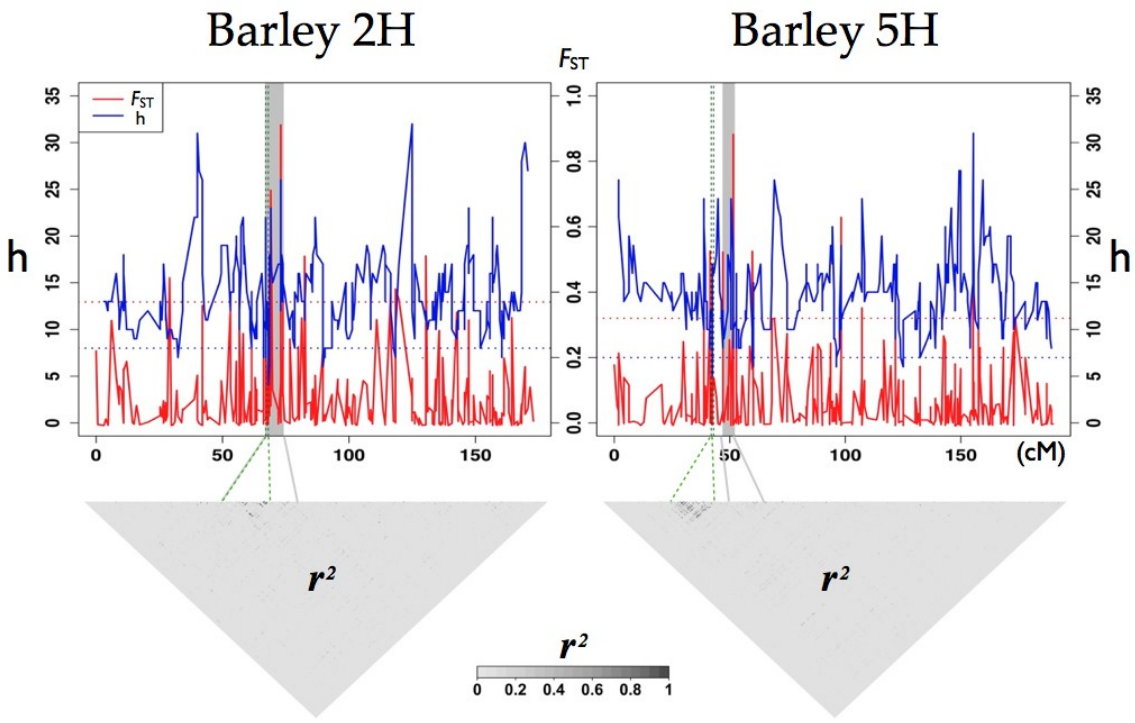


Figure S4.5 Population genetic analysis of the two high *F*ST regions.
Top panel, haplotype number (blue curve) and *F*ST between the Eastern and Western populations (red curve). The number of haplotypes present across linkage group 2H and 5H was calculated in overlapping 5-SNP windows with 1-SNP increments. The high *F*ST regions are marked in grey and the centromeres by green dashed lines. Below, LD (r^2) is plotted across 2H and 5H.

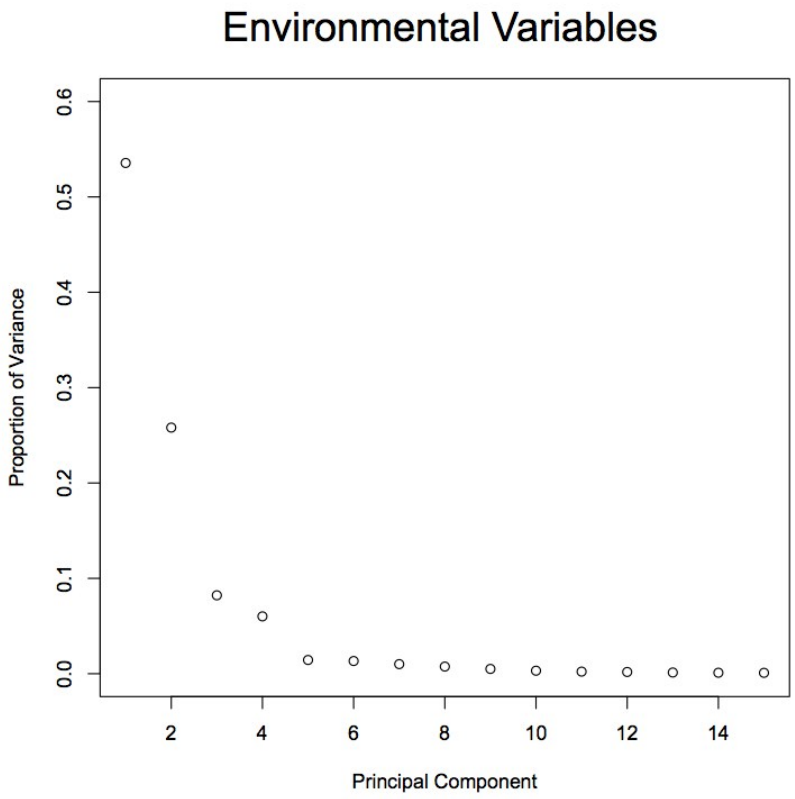
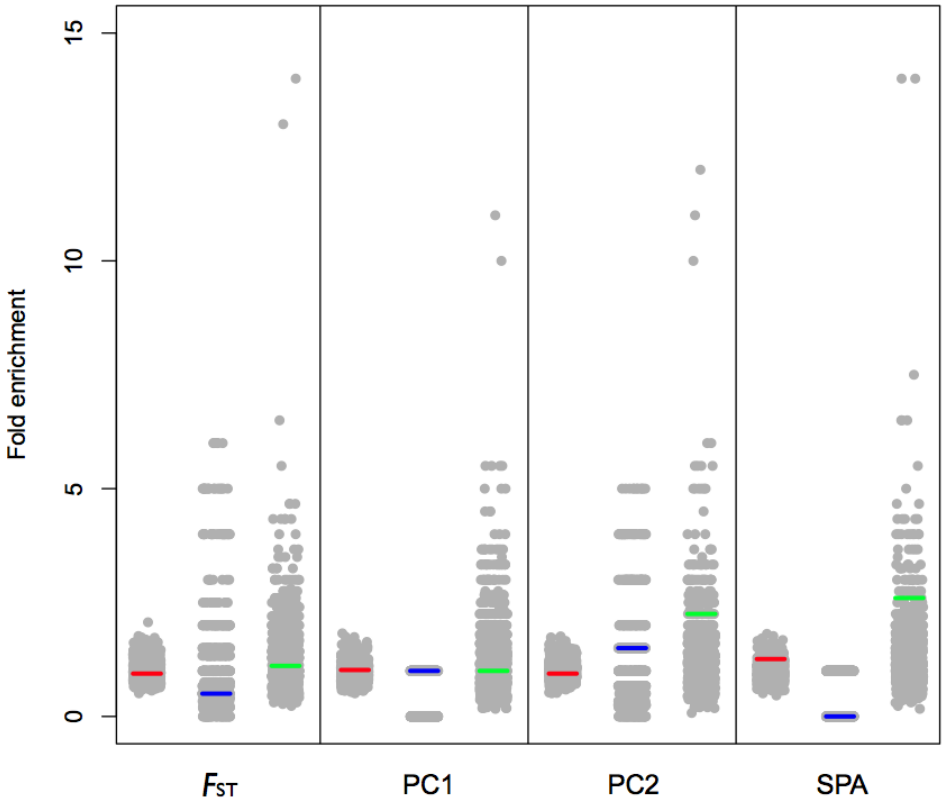


Figure S4.6 The proportion of variance explained by each PC of environmental variables.

(A)



(B)

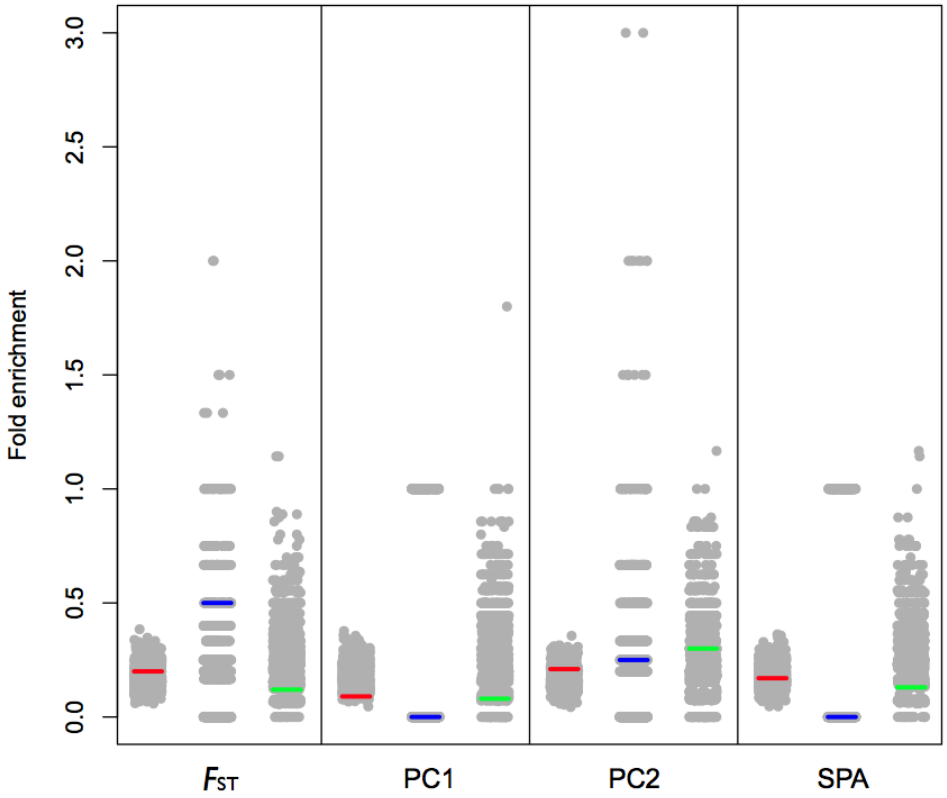


Figure S4.7 Enrichment analysis for (A) genic versus non-genic and (B) nonsynonymous versus synonymous SNPs.
Enrichment of candidates was done by resampling the number of SNPs in each candidate list (outliers in F_{ST}, environmental association or SPA analyses) randomly from the genome 1000 times. Enrichment analysis was done for all SNPs genome-wide, centromere regions and the two putative translocations or inversions. The observed values from the data is indicated by the colored horizontal line, red: genome-wide; blue: centromere regions; green: the two putative translocations or inversions.

4.5.2 Supplementary Tables

Table S4.1 The name, repeat length, repeat unit length, total size and heterozygosity of the 29 microsatellites used in this study.

Microsatellite	# Repeat	Repeat unit length	Total size (bp)	Observed Heterozygosity
Bmag905	14	2	178-228	0.01
Bmag006	17	2	105-223	0.10
Bmac129	28	2	132-204	0.00
Bmac67	18	2	130-262	0.04
Bmag749	11	2	93-199	0.01
Bmac134	28	2	117-189	0.01
Bmac156	(AC)22(AT)5	2	103-199	0.01
Bmac213	23	2	131-213	0.09
Bmac316	19	2	138-234	0.01
Bmag369	16	2	196-216	0.00
Bmag718	(GA)18(AG)6	2	165-217	0.02
Bmag877	15	2	153-289	0.04
EBmac603	10	2	155-257	0.09
HVM06	9	2	118-202	0.23
HVMLOH1A	6	2	147-203	0.00
GMS1	(CT)7TTT(CT)2	2	123-161	0.18
Bmag0496	20	2	139-287	0.03
HVHVA1	5	3	134-140	0.01
Bmac18	11	2	131-145	0.00
Bmag382	(AG)7AA(AG)7	2	103-109	0.00
Scssr02748	12	2	144-158	0.01
Scssr10148	10	2	178-230	0.04
Scssr08447	6	3	172-182	0.00
Scssr05939	5	2	150-160	0.03
Bmag211	16	2	150-198	0.02
Hvltppb	10	2	205-229	0.75
Scssr02306	13	2	150-164	0.00
Scssr15864	4	3	146-184	0.08
Scssr25691	17	2	208-250	0.07

Table S4.2 Environmental variables and abbreviations used in this study.

Environmental variable	Abbreviation
Annual mean temperature	bio1
Mean diurnal range	bio2
Isothermality (bio2/bio7)*100	bio3
Temperature seasonality (standard deviation*100)	bio4
Min temperature of the coldest month	bio6
Mean temperature of the wettest quarter	bio8
Mean temperature of the driest quarter	bio9
Mean temperature of the coldest quarter	bio11
Annual precipitation	bio12
Precipitation of the wettest month	bio13
Precipitation of the driest month	bio14
Precipitation seasonality (coefficient of variation)	bio15
Precipitation of the wettest quarter	bio16
Precipitation of the driest quarter	bio17
Precipitation of the coldest quarter	bio19
Monthly minimum and maximum temperature	tmin#, tmax#
Monthly total precipitation	prec#

Table S4.3 SNPs with F_{ST} based on the Eastern and Western populations above 95th percentile genome-wide

SNP information includes genetic position, GenBank ID, gene short name, in non-coding or coding region (1st, 2nd or 3rd positions), and silent or replacement information.

SNP Name	Chr	cM	F_{ST}	GenBank ID	Gene Short Name	Position	Silent
11_10017	2H	81.31	0.32	AK353943	-	3	yes
11_10056	7H	33.49	0.41	AK364330	-	non-coding	yes
11_10116	5H	41.45	0.36	AK356434	-	3	yes
11_10243	2H	67.08	0.41	XM_003579657	LOC100838855	non-coding	yes
11_10253	3H	102.66	0.36	-	-	-	-
11_10424	4H	56.22	0.42	AK355297	-	3	yes
11_10498	2H	53.09	0.35	AK354555	-	1	yes
11_10536	5H	155.23	0.49	AK354787	-	3	yes
11_10614	4H	111.81	0.33	AK373775	-	3	yes
11_10644	1H	128.63	0.32	AK356375	-	non-coding	yes
11_10685	2H	72.99	0.47	AK362284	-	3	yes
11_10773	7H	81.78	0.36	AK373162	-	non-coding	yes
11_10813	3H	83.58	0.35	-	-	-	-
11_10956	7H	27.69	0.38	AK365588	-	3	yes
11_11024	5H	121.67	0.48	XM_003561753	LOC100839369	3	yes
11_11094	2H	119.72	0.38	-	-	-	-
11_11111	6H	139.09	0.39	AK366470	-	non-coding	yes
11_11243	7H	121.36	0.33	-	-	-	-
11_11432	5H	38.78	0.56	-	-	-	-
11_20029	6H	133.25	0.51	AK354118	-	1	no
11_20086	2H	110.93	0.32	AK358640	-	non-coding	yes
11_20133	1H	132.16	0.47	AK361091	-	non-coding	yes
11_20251	2H	68.56	0.36	AK370757	-	non-coding	yes
11_20283	5H	51.51	0.32	XM_003576993	LOC100831472	3	yes
11_20347	5H	121.67	0.48	AK359986	-	3	yes
11_20390	2H	72.99	0.46	FN179383	SBE2a	3	yes
11_20476	2H	67.08	0.33	BT087333	-	non-coding	yes
11_20482	4H	69.24	0.31	AK355084	-	3	yes
11_20498	2H	116.5	0.4	AK359654	-	3	yes
11_20577	6H	80.06	0.31	AK355031	-	non-coding	yes
11_20620	6H	78.52	0.31	AK366751	-	non-coding	yes
11_20669	2H	68.56	0.34	AK356298	-	non-coding	yes
11_20798	1H	48.99	0.32	AK354088	-	non-coding	yes
11_20904	6H	72.17	0.36	-	-	-	-
11_21302	7H	81.78	0.33	XM_003560861	LOC100844834	non-coding	yes
11_21399	2H	72.99	0.69	AK357878	-	1	no
11_21406	2H	142.67	0.34	AK370573	-	2	no
11_21447	5H	41.45	0.53	AK364513	-	2	no
11_21452	5H	155.23	0.53	AK372156	-	1	no

11_21502	3H	76.43	0.55	AK356987	-	3	yes
11_21504	4H	80.73	0.45	-	-	-	-
12_10053	4H	76.31	0.46	AK370758	-	3	yes
12_10071	6H	130.38	0.47	AK355485	-	non-coding	yes
12_10154	2H	69.05	0.71	AK355324	-	non-coding	yes
12_10170	4H	88.7	0.44	AK373474	-	non-coding	yes
12_10171	4H	43.72	0.34	AK368127	-	1	no
12_10199	6H	49.67	0.64	AK376992	-	non-coding	yes
12_10203	5H	59.72	0.52	AK356265	-	3	yes
12_10219	0	0	0.33	AY039003	Xantha-f	3	yes
12_10264	5H	47.04	0.52	WHTE1A	E1	non-coding	yes
12_10284	0	0	0.47	XM_003564030	LOC100843138	3	yes
12_10347	4H	43.72	0.5	AK362515	-	non-coding	yes
12_10472	2H	147.37	0.31	AK356296	-	non-coding	yes
12_10497	6H	56.06	0.48	XM_003570288	LOC100839823	1	no
12_10543	7H	121.36	0.4	DQ529207	Xantha-h	non-coding	yes
12_10591	6H	59.25	0.44	AK361815	-	1	no
12_10633	5H	69.29	0.32	AK371801	-	non-coding	yes
12_10634	5H	68.21	0.32	AK360310	-	3	yes
12_10689	0	0	0.38	AK369452	-	non-coding	yes
12_10810	4H	37.88	0.69	AK366265	-	non-coding	yes
12_11030	2H	6.09	0.31	AK364311	-	non-coding	yes
12_11107	1H	47.21	0.35	AK364999	-	non-coding	yes
12_11139	4H	111.81	0.38	Y14573	Mlo	3	yes
12_11151	5H	51.51	0.88	AK354730	-	3	yes
12_11184	7H	118.44	0.42	AK358083	-	non-coding	yes
12_11269	0	0	0.41	AK361959	-	1	no
12_11271	1H	136.7	0.4	-	-	-	-
12_11288	2H	67.08	0.52	AK366035	-	3	no
12_11310	3H	13.13	0.36	AK372013	-	3	yes
12_11316	2H	73.89	0.49	AK354712	-	non-coding	yes
12_11324	2H	72.99	0.91	AK356277	-	non-coding	yes
12_11377	7H	84.3	0.34	AK361273	-	3	yes
12_11408	0	0	0.41	-	-	-	-
12_20196	2H	67.08	0.55	AK362518	-	non-coding	yes
12_20235	2H	61.49	0.4	AK376421	-	non-coding	yes
12_20278	5H	59.72	0.52	AK356265	-	1	no
12_20326	2H	42.01	0.36	AK365480	-	non-coding	yes
12_20593	2H	29.05	0.44	AK367163	-	non-coding	yes
12_20760	4H	138.7	0.39	-	-	-	-
12_20981	5H	51.51	0.85	AK370568	-	3	yes
12_20989	2H	130.38	0.51	XM_003580352	LOC100845048	2	no
12_21003	0	0	0.76	-	-	-	-
12_21117	4H	0	0.32	AK359776	-	non-coding	yes
12_21234	7H	68.89	0.54	AK356095	-	2	no

12_21319	7H	82.41	0.37	XM_003573274	LOC100845308	non-coding	yes
12_30004	7H	81.78	0.34	-	-	-	-
12_30049	2H	118.39	0.41	-	-	-	-
12_30056	5H	107.19	0.35	AK356430	-	non-coding	yes
12_30068	2H	67.08	0.52	AK364966	-	3	yes
12_30164	7H	118.44	0.41	AK360970	-	non-coding	yes
12_30250	3H	106.67	0.65	AK356601	-	non-coding	yes
12_30441	6H	58.48	0.36	XM_003570517	LOC100832867	3	yes
12_30492	7H	81.78	0.41	EU961304	-	1	no
12_30504	5H	173.5	0.32	-	-	-	-
12_30563	7H	81.78	0.33	AJ582181	core3ft	2	no
12_30581	7H	79.08	0.45	AK356791	-	3	yes
12_30600	7H	81.78	0.54	AK362488	-	non-coding	yes
12_30616	3H	78.25	0.58	AK365474	-	non-coding	yes
12_30694	1H	45.2	0.6	-	-	-	-
12_30737	3H	59.83	0.39	AK361759	-	3	yes
12_30765	6H	59.25	0.44	AK356029	-	3	yes
12_30823	2H	164.35	0.32	AK375658	-	3	no
12_30850	5H	98.2	0.63	DQ480160	CBF4B	non-coding	yes
12_30956	6H	142.2	0.35	-	-	-	-
12_30988	4H	111.81	0.52	Y14573	Mlo	non-coding	yes
12_31017	3H	67.86	0.31	AK356796	-	non-coding	yes
12_31021	2H	82.44	0.51	AF112963	Cht2	non-coding	yes
12_31032	5H	52.86	0.33	AF326715	adh3	non-coding	yes
12_31035	5H	52.86	0.33	DQ195967	adh3	non-coding	yes
12_31043	6H	112.39	0.31	AY349220	Dhn5	3	yes
12_31062	5H	51.51	0.48	AK365941	-	3	yes
12_31064	5H	51.51	0.68	AK365941	-	3	yes
12_31100	2H	142.67	0.34	AK370573	-	2	no
12_31202	0	0	0.33	AK358464	-	non-coding	yes
12_31203	0	0	0.34	AK368679	-	non-coding	yes
12_31242	3H	82.62	0.36	AK369382	-	non-coding	yes
12_31246	4H	92.41	0.48	AK362249	-	non-coding	yes
12_31274	6H	52.85	0.38	AK357603	-	non-coding	yes
12_31385	4H	77.66	0.32	AK357832	-	non-coding	yes

Table S4.4 Environmental variable and the corresponding loadings for the first two principal components.

PC1		PC2	
Variable	Loadings	Variable	Loadings
bio11	-0.223	bio12	-0.316
tmin2	-0.222	bio16	-0.314
tmax1	-0.222	bio13	-0.310
tmax2	-0.221	prec2	-0.306
tmax12	-0.221	bio19	-0.303
tmax11	-0.220	prec1	-0.297
bio6	-0.220	prec12	-0.295
tmin1	-0.220	prec11	-0.292
tmin3	-0.220	prec3	-0.286
tmin12	-0.218	prec10	-0.180
tmin11	-0.215	prec4	-0.128
tmin10	-0.215	alt	-0.089
tmax3	-0.211	bio15	-0.058
bio1	-0.209	tmin11	-0.039
tmax10	-0.201	tmin12	-0.038
tmin4	-0.196	tmin1	-0.031
tmax4	-0.179	bio6	-0.030
bio3	-0.178	tmin10	-0.019
bio15	-0.176	tmin2	-0.013
bio9	-0.116	bio2	0.199
bio8	-0.111	tmax4	0.147
prec12	-0.080	bio8	0.091
prec1	-0.079	bio4	0.088
bio19	-0.068	tmax3	0.086
prec11	-0.058	bio9	0.086
bio13	-0.052	tmin4	0.071
bio16	-0.048	tmax10	0.068
prec2	-0.047	bio1	0.065
bio12	0.003	tmax2	0.047
prec3	0.020	bio3	0.041
bio2	0.026	tmax11	0.027
prec10	0.091	tmax1	0.023
bio14	0.112	tmax12	0.022
bio17	0.120	tmin3	0.018
prec4	0.141	bio14	0.012
alt	0.154	bio17	0.004
bio4	0.162	bio11	0.003

Table S4.5 SNPs with Bayes Factor from environmental association analysis (Bayenv) above 95th percentile genome-wide

SNP information includes genetic position, GenBank ID, gene short name, in non-coding or coding region (1st, 2nd or 3rd positions), and silent or replacement information. (A) PC1; (B) PC2.

(A)

SNP Name	Chr	cM	Bayes Factor	GenBank ID	Gene Short Name	Position	Silent
11_10011	3H	67.86	9.22	AK371824	-	non-coding	yes
11_10090	4H	86.01	3.91	AK376851	-	non-coding	yes
11_10172	3H	82.62	5.04	AK366339	-	non-coding	yes
11_10184	3H	110.75	3.72	AK359211	-	3	yes
11_10208	4H	2.6	3.76	AK359180	-	3	yes
11_10232	7H	29.05	3.89	-	-	-	-
11_10247	4H	87.01	4.32	AJ585385	bcb	3	no
11_10253	3H	102.66	6.72	-	-	-	-
11_10446	2H	140.69	4.00	XM_003560174	LOC100837523	1	no
11_10641	5H	59.72	4.10	AK355250	-	3	yes
11_10653	3H	69.4	4.18	AK356876	-	2	no
11_10744	1H	23.62	4.80	AK353610	-	non-coding	yes
11_10813	3H	83.58	8.40	-	-	-	-
11_10818	2H	90.48	5.18	AK360545	-	non-coding	yes
11_10925	3H	66.62	3.83	AK363708	-	non-coding	yes
11_10989	2H	125.97	6.42	BLYCLDAB	cold-regulated	non-coding	yes
11_11111	6H	139.09	4.38	AK366470	-	non-coding	yes
11_11211	2H	67.08	4.06	-	-	-	-
11_11361	0	0	9.94	JF796668	CBF4	2	no
11_11497	5H	151.35	6.31	AK358288	-	3	yes
11_11521	7H	127.74	18.19	AK376764	-	3	yes
11_20012	4H	43.72	3.87	AK374823	-	non-coding	yes
11_20078	5H	155.66	5.77	AK376547	-	3	yes
11_20113	7H	49.61	5.40	AK355306	-	non-coding	yes
11_20133	1H	132.16	3.67	AK361091	-	non-coding	yes
11_20251	2H	68.56	5.27	AK370757	-	non-coding	yes
11_20306	5H	50.53	5.22	AK365203	-	non-coding	yes
11_20495	7H	18.73	3.96	AK359539	-	non-coding	yes
11_20537	6H	142.2	4.60	FJ853600	GSL2	non-coding	yes
11_20583	3H	66.62	4.92	AK365999	-	non-coding	yes
11_20620	6H	78.52	7.50	AK366751	-	non-coding	yes
11_20626	3H	109.43	3.65	AK366712	-	3	yes
11_20722	7H	13.7	5.24	AK377052	-	3	yes
11_20797	3H	5.36	3.96	AK375179	-	non-coding	yes
11_21030	6H	44.96	3.79	AK375921	-	2	no
11_21083	3H	111.49	4.11	AK372580	-	non-coding	yes
11_21174	1H	8.96	4.49	-	-	-	-

11_21191	4H	70.77	3.46	AK355907	-	non-coding	yes
11_21192	1H	89.77	5.99	AK357712	-	non-coding	yes
11_21193	1H	42.42	7.48	AK363338	-	non-coding	yes
11_21201	7H	98.95	5.56	-	-	-	-
11_21216	6H	59.25	5.25	AK376749	-	3	yes
11_21274	2H	154.39	3.85	AK354727	-	non-coding	yes
11_21296	4H	71.71	4.88	AK353992	-	non-coding	yes
11_21325	5H	123.12	3.84	AK362100	-	non-coding	yes
11_21399	2H	72.99	8.21	AK357878	-	1	no
11_21406	2H	142.67	3.68	AK370573	-	2	no
11_21459	2H	143.18	4.78	AM039897	ahh1	2	no
11_21502	3H	76.43	3.57	AK356987	-	3	yes
12_10122	3H	143.33	4.09	AK369540	-	non-coding	yes
12_10154	2H	69.05	14.43	AK355324	-	non-coding	yes
12_10199	6H	49.67	4.42	AK376992	-	non-coding	yes
12_10268	7H	81.78	5.32	XM_003574799	LOC100825330	3	yes
12_10535	1H	91.73	9.41	AK358527	-	3	yes
12_10678	3H	71.26	4.44	AK356724	-	non-coding	yes
12_10810	4H	37.88	3.49	AK366265	-	non-coding	yes
12_10824	4H	102.93	4.39	XM_003577507	LOC100843401	non-coding	yes
12_10938	1H	44.59	4.96	AK359548	-	1	no
12_11030	2H	6.09	4.18	AK364311	-	non-coding	yes
12_11051	7H	105.73	7.91	-	-	-	-
12_11151	5H	51.51	3.94	AK354730	-	3	yes
12_11154	3H	147.57	3.85	-	-	-	-
12_11271	1H	136.7	11.25	-	-	-	-
12_11288	2H	67.08	5.01	AK366035	-	3	no
12_11324	2H	72.99	41.73	AK356277	-	non-coding	yes
12_11386	0	0	5.77	-	-	-	-
12_11408	0	0	4.50	-	-	-	-
12_11455	6H	44.96	4.34	AK362081	-	non-coding	yes
12_11494	6H	111.74	3.53	AK368552	-	non-coding	yes
12_20235	2H	61.49	5.61	AK376421	-	non-coding	yes
12_20760	4H	138.7	8.63	-	-	-	-
12_20867	5H	167.4	3.85	AK359088	-	non-coding	yes
12_20981	5H	51.51	4.68	AK370568	-	3	yes
12_21117	4H	0	5.34	AK359776	-	non-coding	yes
12_21234	7H	68.89	3.83	AK356095	-	2	no
12_21319	7H	82.41	4.40	XM_003573274	LOC100845308	non-coding	yes
12_21337	2H	67.08	3.88	AK358015	-	3	yes
12_30004	7H	81.78	4.04	-	-	-	-
12_30060	4H	62.81	9.76	AK365085	-	non-coding	yes
12_30064	3H	50.19	3.70	AK367843	-	non-coding	yes
12_30068	2H	67.08	5.01	AK364966	-	3	yes
12_30080	5H	60.21	3.66	-	-	-	-

12_30135	3H	179.81	3.94	AK356358	-	non-coding	yes
12_30170	3H	92.73	3.50	AK353673	-	3	yes
12_30206	2H	67.08	3.88	-	-	-	-
12_30367	3H	149.45	15.26	AK353825	-	non-coding	yes
12_30404	1H	42.42	4.06	AK355367	-	3	yes
12_30475	7H	82.41	5.64	AK372209	-	non-coding	yes
12_30491	2H	49.5	3.90	-	-	-	-
12_30492	7H	81.78	5.94	EU961304	-	1	no
12_30504	5H	173.5	6.24	-	-	-	-
12_30554	4H	102.93	5.72	AK354013	-	2	no
12_30574	7H	82.41	4.96	AK370386	-	3	yes
12_30581	7H	79.08	18.82	AK356791	-	3	yes
12_30674	2H	85.52	4.34	AK369978	-	non-coding	yes
12_30715	1H	3.21	4.57	AK362579	-	non-coding	yes
12_30724	2H	72.99	3.51	AK368064	-	3	yes
12_30743	3H	87.8	4.49	AK364775	-	non-coding	yes
12_30745	5H	55.44	3.61	AK366468	-	3	yes
12_30779	0	0	5.07	AK376832	-	non-coding	yes
12_30781	2H	9.12	4.05	AK361785	-	non-coding	yes
12_30782	6H	53.54	6.44	-	-	-	-
12_30806	7H	99.94	4.96	-	-	-	-
12_31032	5H	52.86	3.59	AF326715	adh3	non-coding	yes
12_31066	0	0	4.85	AK372734	-	3	yes
12_31081	1H	143.2	21.95	AK356376	-	3	yes
12_31100	2H	142.67	3.68	AK370573	-	2	no
12_31218	2H	67.08	5.54	AK368932	-	2	no
12_31234	5H	141.88	9.78	AK358248	-	non-coding	yes
12_31357	0	0	6.10	-	-	-	-
12_31363	7H	116.94	3.45	AK375416	-	non-coding	yes
12_31381	1H	44.59	3.44	-	-	-	-
12_31383	2H	84.96	3.84	AK363632	-	non-coding	yes
12_31385	4H	77.66	4.68	AK357832	-	non-coding	yes
12_31392	6H	127.76	5.64	AK354269	-	non-coding	yes
12_31424	2H	101.72	3.69	AK375713	-	non-coding	yes
12_31486	4H	6.86	30.04	-	-	-	-

(B)

SNP Name	Chr	cM	Bayes Factor	GenBank ID	Gene Short Name	Position	Silent
11_10006	1H	76.92	4.55	AM502852	tub4	3	yes
11_10052	4H	76.31	3.08	AK370758	-	non-coding	yes
11_10129	6H	44.96	4.94	AK370146	-	3	yes
11_10172	3H	82.62	3.72	AK366339	-	non-coding	yes
11_10194	2H	67.08	2.86	AK359415	-	3	yes
11_10217	5H	149.94	3.16	AK354403	-	non-coding	yes

11_10240	5H	42.41	2.81	AK359441	-	non-coding	yes
11_10329	2H	168.26	3.71	AK362809	-	non-coding	yes
11_10424	4H	56.22	2.83	AK355297	-	3	yes
11_10522	1H	101.34	6.60	AK360842	-	1	yes
11_10580	5H	29.9	18.54	AK372891	-	non-coding	yes
11_10641	5H	59.72	5.49	AK355250	-	3	yes
11_10653	3H	69.4	2.96	AK356876	-	2	no
11_10767	3H	179.16	4.15	-	-	-	-
11_10780	2H	137.03	6.67	AK359167	-	3	yes
11_10919	2H	42.01	2.92	AK368475	-	non-coding	yes
11_11019	4H	146.48	2.85	FN179393	BAM1	2	no
11_11111	6H	139.09	4.09	AK366470	-	non-coding	yes
11_11147	6H	95.7	2.92	GU258512	-	non-coding	yes
11_11219	7H	82.2	5.81	AK368527	-	2	no
11_11406	6H	6.54	5.20	AF166121	Big1	3	yes
11_11521	7H	127.74	4.21	AK376764	-	3	yes
11_20114	4H	44.99	3.66	AK360127	-	3	yes
11_20129	5H	42.41	2.90	AK362176	-	non-coding	yes
11_20179	5H	42.41	2.85	AK357883	-	non-coding	yes
11_20260	1H	42.42	7.95	AK355367	-	3	yes
11_20306	5H	50.53	4.87	AK365203	-	non-coding	yes
11_20332	5H	49.16	3.04	WHTE1A	E1	non-coding	yes
11_20347	5H	121.67	4.55	AK359986	-	3	yes
11_20355	6H	120.69	3.68	AK353922	-	non-coding	yes
11_20383	1H	134.96	5.76	XM_003567791	LOC100830416	3	yes
11_20390	2H	72.99	10.41	FN179383	SBE2a	3	yes
11_20438	2H	69.05	4.78	AK364463	-	non-coding	yes
11_20485	7H	91.67	4.67	-	-	-	-
11_20498	2H	116.5	2.96	AK359654	-	3	yes
11_20527	3H	142.17	5.48	XM_003572099	LOC100831350	3	yes
11_20620	6H	78.52	12.93	AK366751	-	non-coding	yes
11_20709	6H	74.65	2.86	AK363507	-	3	yes
11_20736	5H	72.68	3.59	AK372237	-	3	yes
11_20797	3H	5.36	4.23	AK375179	-	non-coding	yes
11_20889	6H	86.54	3.18	XM_002454244	-	non-coding	yes
11_20924	4H	75.44	5.18	AK376504	-	1	no
11_20958	5H	42.41	2.90	AK372310	-	3	yes
11_20996	6H	104.5	5.87	AK360096	-	3	yes
11_21000	1H	45.85	6.15	AK370060	-	3	yes
11_21040	5H	42.41	2.81	AK370183	-	non-coding	yes
11_21181	2H	155.68	6.75	AK354575	-	3	yes
11_21193	1H	42.42	7.31	AK363338	-	non-coding	yes
11_21201	7H	98.95	10.97	-	-	-	-
11_21399	2H	72.99	8.83	AK357878	-	1	no
11_21494	7H	81.78	3.46	-	-	-	-

12_10053	4H	76.31	6.89	AK370758	-	3	yes
12_10063	4H	44.99	3.28	AK360127	-	non-coding	yes
12_10089	7H	91.12	3.90	-	-	-	-
12_10154	2H	69.05	3.05	AK355324	-	non-coding	yes
12_10159	1H	42.42	2.78	FN555319	pdil4-1	1	no
12_10300	1H	42.42	4.35	EU131177	PR-17c	non-coding	yes
12_10313	0	0	4.12	NM_001061695	Os05g0311000	-	-
12_10392	6H	72.17	3.45	AK361836	-	3	yes
12_10657	7H	61.67	5.82	AK366264	-	3	yes
12_10717	2H	82.98	6.74	-	-	-	-
12_10810	4H	37.88	3.63	AK366265	-	non-coding	yes
12_10824	4H	102.93	2.92	XM_003577507	LOC100843401	non-coding	yes
12_10938	1H	44.59	5.96	AK359548	-	1	no
12_11078	0	0	2.81	AK372406	-	non-coding	yes
12_11151	5H	51.51	5.38	AK354730	-	3	yes
12_11271	1H	136.7	6.00	-	-	-	-
12_11324	2H	72.99	23.02	AK356277	-	non-coding	yes
12_11357	1H	30.15	7.86	AK357047	-	1	no
12_11368	2H	160.46	4.06	AK360506	-	non-coding	yes
12_11468	0	0	3.13	AK359839	-	non-coding	yes
12_20235	2H	61.49	6.27	AK376421	-	non-coding	yes
12_20424	0	0	4.52	AK373140	-	non-coding	yes
12_20649	0	0	3.53	-	-	-	-
12_20760	4H	138.7	5.50	-	-	-	-
12_20793	2H	103.13	3.97	AK371934	-	2	no
12_20981	5H	51.51	4.73	AK370568	-	3	yes
12_21003	0	0	3.04	-	-	-	-
12_21131	1H	59.99	3.29	AK362065	-	3	yes
12_21462	5H	143.29	2.83	AF109194	-	non-coding	yes
12_30005	3H	77.37	2.98	-	-	-	-
12_30060	4H	62.81	10.00	AK365085	-	non-coding	yes
12_30170	3H	92.73	2.95	AK353673	-	3	yes
12_30204	1H	89.77	3.69	AK356336	-	3	yes
12_30226	4H	89.36	8.78	-	-	-	-
12_30242	7H	30.69	3.15	XM_003580160	LOC100839055	1	yes
12_30250	3H	106.67	4.57	AK356601	-	non-coding	yes
12_30275	2H	72.99	3.23	-	-	-	-
12_30342	3H	114.36	3.00	-	-	-	-
12_30389	7H	81.78	3.29	FN179370	AGP-S1a	3	yes
12_30403	1H	128.04	6.73	-	-	-	-
12_30404	1H	42.42	10.32	AK355367	-	3	yes
12_30438	1H	42.42	2.78	AK376647	-	3	yes
12_30475	7H	82.41	23.08	AK372209	-	non-coding	yes
12_30491	2H	49.5	11.08	-	-	-	-
12_30494	5H	171.58	8.49	JX046065	ERS1a	3	yes

12_30504	5H	173.5	4.08	-	-	-	-
12_30528	7H	42.26	9.06	-	-	-	-
12_30574	7H	82.41	4.07	AK370386	-	3	yes
12_30581	7H	79.08	7.89	AK356791	-	3	yes
12_30651	6H	7.87	4.28	AK358128	-	non-coding	yes
12_30674	2H	85.52	2.89	AK369978	-	non-coding	yes
12_30768	5H	42.41	2.81	AK371364	-	non-coding	yes
12_30781	2H	9.12	3.22	AK361785	-	non-coding	yes
12_30976	5H	1.91	3.40	SEG_AY643842 S	-	non-coding	yes
12_30988	4H	111.81	3.04	Y14573	Mlo	non-coding	yes
12_30993	4H	49.43	4.61	XM_003577562	LOC100833831	3	yes
12_31032	5H	52.86	8.13	AF326715	adh3	non-coding	yes
12_31035	5H	52.86	9.13	DQ195967	adh3	non-coding	yes
12_31064	5H	51.51	3.07	AK365941	-	3	yes
12_31081	1H	143.2	3.39	AK356376	-	3	yes
12_31183	5H	50.53	3.78	AK362952	-	non-coding	yes
12_31207	0	0	5.27	AK356841	-	3	yes
12_31234	5H	141.88	2.78	AK358248	-	non-coding	yes
12_31270	0	0	3.11	AK361597	-	non-coding	yes
12_31377	1H	128.04	6.56	AK374439	-	3	yes
12_31424	2H	101.72	3.59	AK375713	-	non-coding	yes
12_31486	4H	6.86	9.19	-	-	-	-

Table S4.6 SNPs with SPA score above 95th percentile genome-wide
SNP information includes genetic position, GenBank ID, gene short name, in non-coding or coding region (1st, 2nd or 3rd positions), and silent or replacement information.

SNP Name	Chr	cM	SPA score	GenBank ID	Gene Short Name	Position	Silent
11_10006	1H	76.92	2.83	AM502852	tub4	3	yes
11_10017	2H	81.31	2.96	AK353943	-	3	yes
11_10056	7H	33.49	3.78	AK364330	-	non-coding	yes
11_10104	5H	141.88	2.86	-	-	-	-
11_10169	7H	106.8	3.34	AK367663	-	non-coding	yes
11_10196	2H	89.68	3.14	AK362580	-	3	yes
11_10253	3H	102.66	2.84	-	-	-	-
11_10477	5H	107.19	3.19	AK354978	-	3	yes
11_10518	5H	88.05	2.93	AK372467	-	1	no
11_10536	5H	155.23	3.25	AK354787	-	3	yes
11_10547	7H	160.97	3.48	AK357875	-	3	yes
11_10557	5H	144.6	2.95	AK371672	-	3	yes
11_10580	5H	29.9	3.02	AK372891	-	non-coding	yes
11_10614	4H	111.81	3.80	AK373775	-	3	yes
11_10619	2H	95.58	3.00	AK373516	-	non-coding	yes
11_10653	3H	69.4	3.67	AK356876	-	2	no
11_10668	4H	50.22	3.04	-	-	-	-
11_10685	2H	72.99	4.37	AK362284	-	3	yes
11_10687	7H	139.9	3.32	AK366838	-	3	yes
11_10939	6H	37.29	3.46	AK371919	-	non-coding	yes
11_11012	7H	149.31	2.93	AK360584	-	1	no
11_11111	6H	139.09	2.80	AK366470	-	non-coding	yes
11_11243	7H	121.36	2.79	-	-	-	-
11_11354	2H	68.07	2.76	AK371682	-	non-coding	yes
11_11432	5H	38.78	3.59	-	-	-	-
11_20018	5H	93.66	3.32	AK376008	-	non-coding	yes
11_20029	6H	133.25	3.16	AK354118	-	1	no
11_20086	2H	110.93	2.96	AK358640	-	non-coding	yes
11_20109	4H	29.34	2.97	AK360657	-	3	yes
11_20133	1H	132.16	2.74	AK361091	-	non-coding	yes
11_20145	4H	1.2	2.76	-	-	-	-
11_20260	1H	42.42	2.73	AK355367	-	3	yes
11_20390	2H	72.99	3.47	FN179383	SBE2a	3	yes
11_20485	7H	91.67	2.97	-	-	-	-
11_20620	6H	78.52	2.78	AK366751	-	non-coding	yes
11_21005	2H	54.92	2.76	-	-	-	-
11_21121	5H	58.65	2.93	AK356836	-	3	yes
11_21192	1H	89.77	4.03	AK357712	-	non-coding	yes
11_21244	5H	51.51	4.22	AK359027	-	3	yes

11_21325	5H	123.12	3.06	AK362100	-	non-coding	yes
11_21340	2H	116.5	2.85	-	-	-	-
11_21399	2H	72.99	4.94	AK357878	-	1	no
11_21406	2H	142.67	3.26	AK370573	-	2	no
11_21447	5H	41.45	3.46	AK364513	-	2	no
11_21452	5H	155.23	2.88	AK372156	-	1	no
11_21502	3H	76.43	3.01	AK356987	-	3	yes
12_10014	3H	173.43	3.22	-	-	-	-
12_10089	7H	91.12	2.76	-	-	-	-
12_10154	2H	69.05	4.02	AK355324	-	non-coding	yes
12_10170	4H	88.7	4.17	AK373474	-	non-coding	yes
12_10199	6H	49.67	3.54	AK376992	-	non-coding	yes
12_10203	5H	59.72	3.93	AK356265	-	3	yes
12_10219	0	0	2.91	AY039003	Xantha-f	3	yes
12_10264	5H	47.04	3.42	WHTE1A	E1	non-coding	yes
12_10284	0	0	3.50	XM_003564030	LOC100843138	3	yes
12_10347	4H	43.72	2.81	AK362515	-	non-coding	yes
12_10392	6H	72.17	3.20	AK361836	-	3	yes
12_10581	7H	82.41	3.14	AK375754	-	3	yes
12_10725	5H	52.86	2.95	AK366248	-	non-coding	yes
12_10810	4H	37.88	3.93	AK366265	-	non-coding	yes
12_10824	4H	102.93	3.58	XM_003577507	LOC100843401	non-coding	yes
12_11151	5H	51.51	4.30	AK354730	-	3	yes
12_11271	1H	136.7	2.77	-	-	-	-
12_11288	2H	67.08	3.93	AK366035	-	3	no
12_11310	3H	13.13	3.29	AK372013	-	3	yes
12_11316	2H	73.89	3.35	AK354712	-	non-coding	yes
12_11324	2H	72.99	5.54	AK356277	-	non-coding	yes
12_11408	0	0	2.73	-	-	-	-
12_11444	1H	54.54	2.76	AJ965495	mcb1	non-coding	yes
12_11498	1H	34.45	3.24	-	-	-	-
12_20278	5H	59.72	3.93	AK356265	-	1	no
12_20413	3H	135.43	2.80	AK356460	-	3	yes
12_20649	0	0	2.80	-	-	-	-
12_20685	7H	94.34	3.24	-	-	-	-
12_20981	5H	51.51	4.22	AK370568	-	3	yes
12_21003	0	0	3.57	-	-	-	-
12_21117	4H	0	4.62	AK359776	-	non-coding	yes
12_21131	1H	59.99	2.90	AK362065	-	3	yes
12_21234	7H	68.89	3.77	AK356095	-	2	no
12_21319	7H	82.41	2.92	XM_003573274	LOC100845308	non-coding	yes
12_30046	4H	105.83	2.94	-	-	-	-
12_30068	2H	67.08	3.93	AK364966	-	3	yes

12_30226	4H	89.36	2.89	-	-	-	-
12_30250	3H	106.67	3.75	AK356601	-	non-coding	yes
12_30344	7H	77.02	2.91	AK359310	-	non-coding	yes
12_30404	1H	42.42	2.79	AK355367	-	3	yes
12_30524	5H	116.66	3.39	AK366243	-	non-coding	yes
12_30577	5H	177.9	2.73	-	-	-	-
12_30581	7H	79.08	2.89	AK356791	-	3	yes
12_30637	6H	72.17	4.10	AK368823	-	non-coding	yes
12_30640	3H	108.7	2.78	AK369817	-	non-coding	yes
12_30644	5H	50.53	2.83	AK374038	-	3	yes
12_30724	2H	72.99	2.83	AK368064	-	3	yes
12_30737	3H	59.83	3.50	AK361759	-	3	yes
12_30745	5H	55.44	3.06	AK366468	-	3	yes
12_30834	5H	88.05	3.20	AK372467	-	non-coding	yes
12_30850	5H	98.2	3.28	DQ480160	CBF4B	non-coding	yes
12_30956	6H	142.2	3.20	-	-	-	-
12_30988	4H	111.81	3.36	Y14573	Mlo	non-coding	yes
12_31017	3H	67.86	3.28	AK356796	-	non-coding	yes
12_31021	2H	82.44	2.85	AF112963	Cht2	non-coding	yes
12_31023	5H	4.15	2.80	AK355143	-	non-coding	yes
12_31032	5H	52.86	3.81	AF326715	adh3	non-coding	yes
12_31035	5H	52.86	3.80	DQ195967	adh3	non-coding	yes
12_31043	6H	112.39	2.77	AY349220	Dhn5	3	yes
12_31062	5H	51.51	3.86	AK365941	-	3	yes
12_31064	5H	51.51	4.23	AK365941	-	3	yes
12_31081	1H	143.2	3.60	AK356376	-	3	yes
12_31100	2H	142.67	3.26	AK370573	-	2	no
12_31179	1H	61.39	2.79	AY738115	cbp1	3	yes
12_31189	2H	67.08	2.76	AK364573	-	3	yes
12_31202	0	0	2.73	AK358464	-	non-coding	yes
12_31246	4H	92.41	3.24	AK362249	-	non-coding	yes
12_31411	0	0	2.81	AK372597	-	2	no
12_31486	4H	6.86	3.75	-	-	-	-
12_31511	0	0	3.01	AK360621	-	non-coding	yes
12_31525	3H	134.05	3.32	AK364139	-	2	no

CHAPTER 5

CONCLUSION

In this dissertation, I used population genetic approaches to identify genomic regions that are subject to selection or local adaptation in teosinte, barley and wild barley. I showed in all chapters that bottom-up approach is useful to identify genomic regions under selection. The bottom-up approach, which uses population genetic analyses, has several advantages. First, the phenotype information is not required. We can identify the most important genetic variants and the most obvious signals in the genome without the influence of measuring phenotypes, especially those phenotypes that are hard to measure; second, segregating variation is also not required to identify genes of interest; third, it is fast and requires fewer samples, and finally it works well in species with limited genomic resources, such as barley (Ross-Ibarra *et al.*, 2007).

However, the bottom-up approach has several limitations. The population genetic analysis based bottom-up approach is expected to miss many loci under selection (Teshima *et al.*, 2006) and expected to have high false-positive rates (Narum and Hess, 2011). The outlier-based approach can miss many of the loci under selection, especially when the marker density is low. With the current marker density, the most obvious association usually comes from chromosomal structural variations, because they are large, have high levels of LD and capture more than one adaptive mutation. Therefore, low SNP density does not preclude identification of adaptive variants with the largest effect, such as the chromosomal structural variations reported in Chapter 2 and 4. Moreover, the outlier-based approach has several assumptions (mentioned in Chapter 1) and the population history is usually very complicated. We need to take into consideration of drift, migration, selection and demographic effects when using any

population genetic models, like the coalescent simulations I performed in Chapter 3 and we need to have a more mechanistic understanding of adaptation (Barrett and Hoekstra, 2011; Kirkpatrick and Kern, 2012). Finally, as also mentioned in (Ross-Ibarra *et al.*, 2007), this population genetic approach alone is insufficient for showing that a particular allele is adaptive. We need to test the functional effects of mutations on proteins and genes on phenotypes.

The history of crop wild relatives is much longer than the history of domesticated crops, which have undergone short and rapid change to meet the needs of humans, such as the loss of seed shattering and increase in seed size. Therefore, natural populations of crop wild relatives have the potential to serve as a source of genetic variants that contribute to favorable agronomic traits, including improved disease resistance, and cold or drought tolerance. The inversion identified in Chapter 2 is not present in domesticated maize. The importance of this discovery is that natural populations of relatives of maize, our most important crop, contain an important genetic variant that potentially improves adaptation to hot and dry environments. In Chapter 4, I also identified two putatively adaptive chromosomal genetic variations in wild barley. Those novel genetic variants and functional adaptations found in crop wild progenitors could be integrated into breeding programs for crop improvement and future crops will be more tolerant of harsh environments.

REFERENCES

Albrechtsen, A., I. Moltke, and R. Nielsen, 2010a Natural selection and the distribution of identity-by-descent in the human genome. Genetics 186: 295-308.

Albrechtsen, A., F. C. Nielsen, and R. Nielsen, 2010b Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27: 2534-2547.

Allard, R. W., A. L. Kahler, and B. S. Weir, 1972 The effect of selection on esterase allozymes in a barley population. Genetics 72: 489-503.

Anderson, A. R., A. A. Hoffmann, S. W. Mckechnie, P. A. Umina, and A. R. Weeks, 2005 The latitudinal cline in the *In (3R) Payne* inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. Mol Ecol 14: 851-858.

Anderson, E., 1946 Maize in Mexico a preliminary survey. Ann Missouri Bot Gard 33: 147-247.

Andolfatto, P., J. D. Wall, and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics 153: 1297-1311.

Ayala, D., M. C. Fontaine, A. Cohuet, D. Fontenille, R. Vitalis *et al.*, 2011 Chromosomal inversions, natural selection and adaptation in the malaria vector *Anopheles funestus*. Mol Biol Evol 28: 745-758.

Balanyá, J., J. M. Oller, R. B. Huey, G. W. Gilchrist, and L. Serra, 2006 Global genetic change tracks global climate warming in *Drosophila subobscura*. Science 313: 1773-1775.

Bansal, V., A. Bashir, and V. Bafna, 2007 Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res 17: 219-230.

Barrett, R. D. H., and H. E. Hoekstra, 2011 Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12: 767-780.

Becquet, C., and M. Przeworski, 2007 A new approach to estimate parameters of speciation models with application to apes. Genome Res 17: 1505-1519.

Bengtsson, B. O., and W. F. Bodmer, 1976 On the increase of chromosome mutations under random mating. Theor Popul Biol 9: 260-281.

Blake, V. C., J. G. Kling, P. M. Hayes, and J. L. Jannink, 2012 The Hordeum Toolbox - The Barley Coordinated Agricultural Project genotype and phenotype resource. Plant Genome 5: 81-91.

Boone, C., H. Bussey, and B. J. Andrews, 2007 Exploring genetic interactions and networks with yeast. Nat Rev Genet 8: 437-449.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Bretting, P. K., and M. M. Goodman, 1989 Karyotypic variation in Mesoamerican races of maize and its systematic significance. Econ Bot 43: 107-124.

Briggs, W. H., M. D. McMullen, B. S. Gaut, and J. Doebley, 2007 Linkage mapping of domestication loci in a large maize teosinte backcross resource. Genetics 177: 1915-1928.

Brown, A. H. D., D. Zohary, and E. Nevo, 1978 Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* Koch in Israel. Heredity 41: 49-62.

Buckler, E. S., T. L. Phelps-Durr, C. S. K. Buckler, R. K. Dawe, J. F. Doebley *et al.*, 1999 Meiotic drive of chromosomal knobs reshaped the maize genome. Genetics 153: 415-426.

Burke, M. K., and M. R. Rose, 2009 Experimental evolution with *Drosophila*. Am J Physiol Regul Integr Comp Physiol 296: R1847-R1854.

Burke, M. K., J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose *et al.*, 2010 Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature 467: 587-590.

Burnham, C. R., 1962 *Discussions in cytogenetics*. Burgess Publishing, Minneapolis.

Caldwell, K. S., J. Russell, P. Langridge, and W. Powell, 2006 Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172: 557-567.

Castermans, D., J. R. Vermeesch, J. P. Fryns, J. G. Steyaert, W. J. M. Van de Ven *et al.*, 2007 Identification and characterization of the *TRIP8* and *REEP3* genes on chromosome 10q21.3 as novel candidate genes for autism. Eur J Hum Genet 15: 422-431.

Cattivelli, L., and D. Bartels, 1990 Molecular cloning and characterization of cold-regulated genes in barley. Plant physiol 93: 1504-1510.

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proc. Roy. Soc. London Ser. B 362-379.

Cheng, C., B. J. White, C. Kamdem, K. Mockaitis, C. Costantini *et al.*, 2012 Ecological genomics of *Anopheles gambiae* along a latitudinal cline in Cameroon: a population resequencing approach. Genetics 190: 1417-1432.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen, 2005a Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496-1502.

Clark, R. M., S. Tavare, and J. Doebley, 2005b Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Mol Biol Evol 22: 2304-2312.

Clegg, M. T., R. W. Allard, and A. L. Kahler, 1972 Is the gene the unit of selection? Evidence from two experimental plant populations. Proc Natl Acad Sci USA 69: 2474-2478.

Close, T. J., P. R. Bhat, S. Lonardi, Y. Wu, N. Rostoks *et al.*, 2009 Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10: 582.

Cockram, J., J. White, D. L. Zuluaga, D. Smith, J. Comadran *et al.*, 2010 Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Nat Acad Sci USA 107: 21611-21616.

Condón, F., C. Gustus, D. C. Rasmusson, and K. P. Smith, 2008 Effect of advanced cycle breeding on genetic diversity in barley breeding germplasm. Crop Sci 48: 1027-1036.
Condón, F., D. C. Rasmusson, E. Schiefelbein, G. Velasquez, and K. P. Smith, 2009 Effect of advanced cycle breeding on genetic gain and phenotypic diversity in barley breeding germplasm. Crop Sci 49: 1751-1761.
Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411-1423.
Coyne, J. A., W. Meyers, A. P. Crittenden, and P. Sniegowski, 1993 The fertility effects of pericentric inversions in *Drosophila melanogaster*. Genetics 134: 487-496.
Dawe, R. K., and W. Z. Cande, 1996 Induction of centromeric activity in maize by suppressor of meiotic drive 1. Proc Natl Acad Sci USA 93: 8512-8517.
de la Pena, R. C., K. P. Smith, F. Capettini, G. J. Muehlbauer, M. Gallo-Meagher *et al.*, 1999 Quantitative trait loci associated with resistance to Fusarium head blight and kernel discoloration in barley. Theor Appl Genet 99: 561-569.
Depaulis, F., L. Brazier, and M. Veuille, 1999 Selective sweep at the *Drosophila melanogaster s*uppressor of hairless locus and its association with the *In(2L)t* inversion polymorphism. Genetics 152: 1017-1024.
DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
Dobzhansky, T., 1950 Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. Genetics 35: 288-302.
Doebley, J. F., 1983 The maize and teosinte male inflorescence: a numerical taxonomic study. Ann Missouri Bot Gard 32-70.
Dudley, J. W., and R. J. Lambert, 2004 100 generations of selection for oil and protein in corn. Plant Breed Rev 24: 79-110.
Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman, and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. Proc Nat Acad Sci USA 95: 4441-4446.
Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.
Fang, Z., T. Pyhäjärvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz *et al.*, 2012 Megabase-scale inversion polymorphism in the wild ancestor of maize. Genetics 191: 883-894.
Fang, Z., A. M. Gonzales, M. L. Durbin, K. K. T. Meyer, B. H. Miller *et al.*, 2013a Tracing the geographic origin of *Ipomoea purpurea* (morning glory) in the Southeastern United States. J Hered 104:666-677.
Fang, Z., A. Eule-Nashoba, C. Powers, T. Y. Kono, S. Takuno *et al.*, 2013b Comparative analyses identify the contributions of exotic donors to disease resistance in a barley experimental population. G3: Genes | Genomes | Genetics 3: 1945-1953.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.
Felsenstein, J., 1993 PHYLIP (Phylogeny Inference Package), version 3.6 (distributed by the author). Department of Genome Sciences, University of Washington, Seattle.
Fetch, T. G., B. J. Steffenson, and E. Nevo, 2003 Diversity and sources of multiple disease resistance in *Hordeum spontaneum*. Plant Dis 87: 1439-1448.
Fitch, W. M., and E. Margoliash, 1967 Construction of phylogenetic trees. Science 155: 279-284.
Francia, E., F. Rizza, L. Cattivelli, A. M. Stanca, G. Galiba *et al.*, 2004 Two loci on chromosome 5H determine low-temperature tolerance in a 'Nure'(winter) × 'Tremois'(spring) barley map. Theor Appl Genet 108: 670-680.
Fukunaga, K., J. Hill, Y. Vigouroux, Y. Matsuoka, G. J. Sanchez *et al.*, 2005 Genetic diversity and population structure of teosinte. Genetics 169: 2241-2254.
Gattepaille, L. M., and M. Jakobsson, 2012 Combining markers into haplotypes can improve population structure inference. Genetics 190: 159-174.
Gonzales, A. M., Z. Fang, M. L. Durbin, K. K. T. Meyer, M. T. Clegg *et al.*, 2012 Nucleotide sequence diversity of floral pigment genes in Mexican productions of *Ipomoea purpurea* (morning glory) accord with a neutral model of evolution. J Hered 103: 863-872.
Goodman, S. J., 1997 R_{ST} Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. Mol Ecol 6: 881-885.
Gore, M. A., J. M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz *et al.*, 2009 A first-generation haplotype map of maize. Science 326: 1115-1117.
Goudet, J., 2005 Hierfstat, a package for R to compute and test hierarchical F- statistics. Mol Ecol Notes 5: 184-186.
Graubard, M. A., 1932 Inversion in *Drosophila melanogaster*. Genetics 17: 81-105.
Grossi, M., E. Giorni, F. Rizza, A. M. Stanca, and L. Cattivelli, 1998 Wild and cultivated barleys show differences in the expression pattern of a cold-regulated gene family under different light and temperature conditions. Plant Mol Biol 38: 1061-1069.
Guerrero, R. F., F. Rousset, and M. Kirkpatrick, 2012 Coalescent patterns for chromosomal inversions in divergent populations. Phil. Trans. R. Soc. B. 367: 430-438.
Günther, T., and G. Coop, 2013 Robust identification of local adaptation from allele frequencies. Genetics 195:205-220.
Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler *et al.*, 2009 Whole population, genome-wide mapping of hidden relatedness. Genome Res 19: 318-326.
Haasl, R. J., and B. A. Payseur, 2010 Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. Heredity 106: 158-171.
Hamblin, M. T., T. J. Close, P. R. Bhat, S. Chao, J. G. Kling *et al.*, 2010 Population structure and linkage disequilibrium in US barley germplasm: implications for association mapping. Crop Sci 50: 556-566.

Harlan, H. V., and M. L. Martini, 1929 A composite hybrid mixture. J Am Soc Agron 21: 487-490.
Harris, D. R., 1990 Vavilov's concept of centres of origin of cultivated plants: its genesis and its influence on the study of agricultural origins. Biol J Linn Soc 39: 7-16.
Hijmans, R. J., L. Guarino, M. Cruz, and E. Rojas, 2001 Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genet Resour Newsl 15-19.
Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis, 2005 Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25: 1965-1978.
Hill, W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. Genet Res 8: 269-294.
Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226-231.
Hoffmann, A. A., C. M. Sgro, and A. R. Weeks, 2004 Chromosomal inversion polymorphisms and adaptation. Trends Ecol Evol 19: 482-488.
Hoffmann, A. A., and L. H. Rieseberg, 2008 Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annu Rev Ecol Evol Syst 39: 21-42.
Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42: 961-967.
Hudson, R. R., M. Kreitman, and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.
Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.
Hudson, R. R., 2007 The variance of coalescent time estimates from DNA sequences. J Mol Evol 64: 702-705.
Hufford, M. B., X. Xu, J. Van Heerwaarden, T. Pyhäjärvi, J. M. Chia *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nat Genet 44: 808-811.
Huynh, L. Y., D. L. Maney, and J. W. Thomas, 2011 Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). Heredity 106: 537-546.
Innan, H., and Y. Kim, 2008 Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. Genetics 179: 1713-1720.
Jakobsson, M., and N. A. Rosenberg, 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801-1806.
Jiang, C., P. W. Chee, X. Draye, P. L. Morrell, C. W. Smith *et al.*, 2000 Multilocus interactions restrict gene introgression in interspecific populations of polyploid *Gossypium* (cotton). Evolution 54: 798-814.

Jones, H., F. J. Leigh, I. Mackay, M. A. Bower, L. M. J. Smith *et al.*, 2008 Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the Fertile Crescent. Mol Biol Evol 25: 2211-2219.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.

Kato Y., T. A., 1975 Cytological studies of maize and teosinte in relation to their origin and evolution. Available at http://www.maizegdb.org/cooperators.php. Accessed November 2013.

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513-1524.

Kirkpatrick, M., and N. Barton, 2006 Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.

Kirkpatrick, M., and A. Kern, 2012 Where's the money? Inversions, genes, and the hunt for genomic targets of selection. Genetics 190: 1153-1155.

Konishi, T., and I. Linde-Laursen, 1988 Spontaneous chromosomal rearrangements in cultivated and wild barleys. Theor Appl Genet 75: 237-243.

Kono, T. Y., K. Seth, J. A. Poland, and P. L. Morrell, 2013 SNPMeta: SNP annotation and SNP metadata collection without a reference genome. Mol Ecol Resour doi: 10.1111/1755-0998.12183.

Kump, K. L., P. J. Bradbury, R. J. Wisser, E. S. Buckler, A. R. Belcher *et al.*, 2011 Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43: 163-168.

Lamb, J. C., J. M. Meyer, and J. A. Birchler, 2007 A hemicentric inversion in the maize line knobless Tama flint created two sites of centromeric elements and moved the kinetochore-forming region. Chromosoma 116: 237-247.

Lande, R., 1984 The expected fixation rate of chromosomal inversions. Evolution 38: 743-752.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359.

Levitan, M., 2001 Studies of linkage in populations. XIV. Historical changes in frequencies of gene arrangements and arrangement combinations in natural populations of *Drosophila robusta*. Evolution 55: 2359-2362.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49-67.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079.

Li, J. Z., T. G. Sjakste, M. S. Röder, and M. W. Ganal, 2003 Development and genetic mapping of 127 new microsatellite markers in barley. Theor Appl Genet 107: 1021-1027.

Li, X., C. N. Topp, and R. K. Dawe, 2010 Maize antibody procedures: immunolocalization and chromatin immunoprecipitation (ChIP), pp. 271-286 in

Plant cytogenetics, genome structure and chromosome function, edited by H. W. Bass, and J. A. Birchler. Springer-Verlag, Berlin/Heidelberg, Germany/New York.
Lin, J. Z., A. H. Brown, and M. T. Clegg, 2001 Heterogeneous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). Proc Natl Acad Sci USA 98: 531-536.
Lin, Y. R., K. F. Schertz, and A. H. Paterson, 1995 Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics 141: 391-411.
Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9: 720-731.
Long, Q., F. A. Rabanal, D. Meng, C. D. Huber, and A. Farlow *et al.*, 2013 Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet 45: 884-890.
Lorenz, A. J., K. P. Smith, and J. L. Jannink, 2012 Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci 52: 1609-1621.
Lowry, D. B., and J. H. Willis, 2010 A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biol 8: e1000500.
Ma, Z., B. J. Steffenson, L. K. Prom, and N. L. V. Lapitan, 2000 Mapping of quantitative trait loci for Fusarium head blight resistance in barley. Phytopathology 90: 1079-1088.
Machado, C. A., T. S. Haselkorn, and M. A. F. Noor, 2007 Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis*. Genetics 175: 1289-1306.
Mackay, T. F., E. A. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10: 565-577.
Maguire, M. P., 1966 The relationship of crossing over to chromosome synapsis in a short paracentric inversion. Genetics 53: 1071-1077.
Maguire, M. P., and R. W. Riess, 1994 The relationship of homologous synapsis and crossing over in a maize inversion. Genetics 137: 281-288.
Mano, Y., F. Omori, and K. Takeda, 2012 Construction of intraspecific linkage maps, detection of a chromosome inversion, and mapping of QTL for constitutive root aerenchyma formation in the teosinte *Zea nicaraguensis*. Mol Breed 29: 137-146.
Martin, J. M., T. K. Blake, and E. A. Hockett, 1991 Diversity among North American spring barley cultivars based on coefficients of parentage. Crop Sci 31: 1131-1137.
Massman, J., B. Cooper, R. Horsley, S. Neate, R. Dill-Macky *et al.*, 2011 Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. Mol Breed 27: 439-454.

Matsumoto, T., T. Tanaka, H. Sakai, N. Amano, H. Kanamori *et al.*, 2011 Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. Plant Physiol 156: 20-28.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, G. Sanchez, E. S. Buckler *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc Natl Acad Sci USA 99: 6080-6084.

Matthews, D. E., V. L. Carollo, G. R. Lazo, and O. D. Anderson, 2003 GrainGenes, the genome database for small-grain crops. Nucleic Acids Res 31: 183-186.

Mayer, K. F., M. Martis, P. E. Hedley, H. Simkova, H. Liu *et al.*, 2011 Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23: 1249-1263.

Mayer, K. F., R. Waugh, J. W. Brown, A. Schulman, P. Langridge *et al.*, 2012 A physical, genetic and functional sequence assembly of the barley genome. Nature 491: 711-716.

McClintock, B., 1931 Cytological observations of deficiencies involving known genes, translocations and an inversion in *Zea mays*. Mo Agric Expt Sta Bull 163: 1-30.

McClintock, B., 1933 The association of non-homologous parts of chromosomes in the mid-prophase of meiosis in *Zea mays*. Z. Zellforsch Mikrosk Anat 19: 191-237.

McClintock, B., 1960 Chromosome constitutions of Mexican and Guatemalan races of maize. Carnegie Inst. Wash. Yearbook 59: 461-472.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351: 652-654.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.

McMullen, M., R. Jones, and D. Gallenberg, 1997 Scab of wheat and barley: a re-emerging disease of devastating impact. Plant Dis 81: 1340-1348.

McVean, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics 175: 1395-1406.

Mesfin, A., K. P. Smith, R. Waugh, R. Dill-Macky, C. K. Evans *et al.*, 2003 Quantitative trait loci for Fusarium head blight resistance in barley detected in a two-rowed by six-rowed population. Crop Sci 43: 307-318.

Moeller, D. A., M. I. Tenaillon, and P. Tiffin, 2007 Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays* ssp. *parviglumis*). Genetics 176: 1799-1809.

Morgan, D. T., 1950 A cytogenetic study of inversions in *Zea mays*. Genetics 35: 153-174.

Morrell, P. L., K. E. Lundy, and M. T. Clegg, 2003 Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. Proc Natl Acad Sci USA 100: 10812-10817.

Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg, 2005 Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. Proc Natl Acad Sci USA 102: 2442-2447.

Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg, 2006 Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics 173: 1705-1723.
Morrell, P. L., and M. T. Clegg, 2007 Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. Proc Natl Acad Sci USA 104: 3289-3294.
Morrell, P. L., A. M. Gonzales, K. K. T. Meyer, and M. T. Clegg, in press Resequencing data indicates a modest effect of domestication on diversity in barley: a cultigen with multiple origins. J Hered 10.1093/jhered/est083
Muñoz-Amatriaín, M., M. J. Moscou, and P. R. Bhat, 2011 An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. Plant Genome 4: 238-249.
Munte, A., J. Rozas, M. Aguade, and C. Segarra, 2005 Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. Genetics 169: 1573-1581.
Narum, S. R., and J. E. Hess, 2011 Comparison of F_{ST} outlier tests for SNP loci under selection. Mol Ecol Resour 11: 184-194.
Nei, M., and W. H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76: 5269-5273.
Nei, M., and T. Maruyama, 1975 Lewontin-Krakauer test for neutral genes. Genetics 80: 395.
Nevo, E., A. Beiles, and D. Zohary, 1986 Genetic resources of wild barley in the Near East: structure, evolution and application in breeding. Biol J Linn Soc 27: 355-380.
Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566-1575.
Nóbrega, C., M. Khadem, M. Aguadé, and C. Segarra, 2008 Genetic exchange versus genetic differentiation in a medium-sized inversion of *Drosophila*: the A2/AST arrangements of *Drosophila subobscura*. Mol Biol Evol 25: 1534-1543.
Orozco- Terwengel, P., M. Kapun, V. Nolte, R. Kofler, T. Flatt *et al.*, 2012 Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Mol Ecol 21: 4931-4941.
Paradis, E., J. Claude, and K. Strimmer, 2004 APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.
Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood *et al.*, 2009 The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551-556.
Pfaff, C. L., E. J. Parra, C. Bonilla, K. Hiester, P. M. McKeigue *et al.*, 2001 Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68: 198-207.
Pinhasi, R., J. Fort, and A. J. Ammerman, 2005 Tracing the origin and spread of agriculture in Europe. PLoS Biol 3: e410.
Pinhasi, R., M. G. Thomas, M. Hofreiter, M. Currat, and J. Burger, 2012 The genetic history of Europeans. Trends Genet 28: 496-505.

Piperno, D. R., A. J. Ranere, I. Holst, J. Iriarte, and R. Dickau, 2009 Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. Proc Natl Acad Sci USA 106: 5019-5024.
Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
R Development Core Team, 2011 R: A language and environment for statistical computing, version 2.12.2. R Foundation for Statistical Computing, Vienna. http://www.r-project.org.
Ramage, R. T., and C. A. Suneson, 1961 Translocation-gene linkages on barley chromosome 7. Crop Sci 1: 319-320.
Ramsay, L., M. Macaulay, S. Degli Ivanissevich, K. MacLean, L. Cardle *et al.*, 2000 A simple sequence repeat-based linkage map of barley. Genetics 156: 1997-2005.
Rasmusson, D. C., and R. L. Phillips, 1997 Plant breeding progress and genetic diversity from de novo variation and elevated epistasis. Crop Sci 37: 303-310.
Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98: 11479-11484.
Rieseberg, L. H., B. Sinervo, C. R. Linder, M. C. Ungerer, and D. M. Arias, 1996 Role of gene interactions in hybrid speciation: evidence from ancient and experimental hybrids. Science 272: 741-745.
Rosenberg, N. A., L. M. Li, R. Ward, and J. K. Pritchard, 2003 Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402-1422.
Ross-Ibarra, J., P. L. Morrell, and B. S. Gaut, 2007 Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci USA 104: 8641-8648.
Ross-Ibarra, J., M. Tenaillon, and B. S. Gaut, 2009 Historical divergence and gene flow in the genus *Zea*. Genetics 181: 1399-1413.
Rostoks, N., S. Mudie, L. Cardle, J. Russell, L. Ramsay *et al.*, 2005 Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics 274: 515-527.
Russell, J., I. K. Dawson, A. J. Flavell, B. Steffenson, E. Weltzien *et al.*, 2011 Analysis of >1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome- level differences in diversity around domestication genes. New Phytol 191: 564-578.
Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.
Saisho, D., and M. D. Purugganan, 2007 Molecular phylogeography of domesticated barley traces expansion of agriculture in the Old World. Genetics 177: 1765-1776.
Sambrook, J., E. Fritsch, and T. Maniatis, 1989 *Molecular cloning: A laboratory manual*, Ed. 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Santos, M., W. Céspedes, J. Balanyà, V. Trotta, F. C. F. Calboli *et al.*, 2005 Temperature- related genetic changes in laboratory populations of *Drosophila*

subobscura: evidence against simple climatic- based explanations for latitudinal clines. Am Nat 165: 258-273.
Sato, K., T. Shin-I, M. Seki, K. Shinozaki, H. Yoshida *et al.*, 2009a Development of 5006 full-length cDNAs in barley: a tool for accessing cereal genomics resources. DNA Res 16: 81-89.
Sato, K., N. Nankaku, and K. Takeda, 2009b A high-density transcript linkage map of barley derived from a single population. Heredity 103: 110-117.
Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.
Schmalenbach, I., J. Léon, and K. Pillen, 2009 Identification and verification of QTLs for agronomic traits using wild barley introgression lines. Theor Appl Genet 118: 483-497.
Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.*, 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112-1115.
Schoen, D. J., and A. H. D. Brown, 2001 The Conservation of Wild Plant Species in Seed Banks. Bioscience 51: 960-966.
Shin, J. H., S. Blay, B. McNeney, and J. Graham, 2006 LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. J Stat Soft 16: 1-10.
Shriner, D., 2011 Investigating population stratification and admixture using eigenanalysis of dense genotypes. Heredity 107: 413-420.
Sim, S. C., M. D. Robbins, A. Van Deynze, A. P. Michel, and D. M. Francis, 2010 Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum L.*). Heredity 106: 927-935.
Slatkin, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics 139: 457-462.
Smith, K. P., D. C. Rasmusson, E. Schiefelbein, J. J. Wiersma, J. V. Wiersma *et al.*, 2010 Registration of 'Rasmusson' barley. J Plant Reg 4: 167-170.
Smith, K. P., A. Budde, R. Dill-Macky, D. C. Rasmusson, E. Schiefelbein *et al.*, 2013 Registration of 'Quest' spring malting barley with improved resistance to Fusarium head blight. J Plant Reg 7: 125-129.
Steffenson, B. J., P. Olivera, J. K. Roy, Y. Jin, K. P. Smith *et al.*, 2007 A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. Aust J Agric Res 58: 532-544.
Stevison, L. S., K. B. Hoehn, and M. A. F. Noor, 2011 Effects of inversions on within- and between-species recombination and divergence. Genome Biol. Evol. 3: 830-841.
Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440-9445.
Suneson, C. A., 1956 An evolutionary plant breeding method. Agron J 48: 188-191.
Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg, 2008 ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics 24: 2498-2504.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.
Takeda, S., and M. Matsuoka, 2008 Genetic approaches to crop improvement: Responding to environmental and population changes. Nat Rev Genet 9: 444-457.
Tenaillon, M. I., J. U'Ren, O. Tenaillon, and B. S. Gaut, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21: 1214-1225.
Tenaillon, O., A. Rodríguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett *et al.*, 2012 The molecular diversity of adaptive convergence. Science 335: 457-461.
Teotónio, H., I. M. Chelo, M. Bradi&cacute, M. R. Rose, and A. D. Long, 2009 Experimental evolution reveals natural selection on standing genetic variation. Nat Genet 41: 251-257.
Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702-712.
Thomas, J. W., M. Cáceres, J. J. Lowman, C. B. Morehouse, M. E. Short *et al.*, 2008 The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. Genetics 179: 1455-1468.
Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman, 2000 Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci USA 97: 7360-7365.
Thornton, K., 2003 Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics 19: 2325-2327.
Tian, F., N. M. Stevens, and E. S. Buckler, 2009 Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc Natl Acad Sci USA 106: 9979-9986.
Ting, Y. C., 1965 Spontaneous chromosome inversions of Guatemalan teosintes (*Zea mexicana*). Genetica 36: 229-242.
Ting, Y. C., 1967 Common inversion in maize and teosinte. Am Nat 101: 87-89.
Ting, Y. C., 1976 Chromosome polymorphism of teosinte. Genetics 83: 737-742.
Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genet 7: e1001336.
Umina, P. A., A. R. Weeks, M. R. Kearney, S. W. McKechnie, and A. A. Hoffmann, 2005 A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. Science 308: 691-693.
Valdes, A. M., M. Slatkin, and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133: 737-749.
van Heerwaarden, J., J. Ross-Ibarra, J. Doebley, J. C. Glaubitz, J. Gonzalez Jde *et al.*, 2010 Fine scale genetic structure in the wild ancestor of maize (*Zea mays* ssp. *parviglumis*). Mol Ecol 19: 1162-1173.

van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman *et al.*, 2011 Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci USA 108: 1088-1092.

van Heerwaarden, J., M. B. Hufford, and J. Ross-Ibarra, 2012 Historical genomics of North American maize. Proc Natl Acad Sci USA 109: 12420-12425.

Varshney, R. K., T. C. Marcel, L. Ramsay, J. Russell, M. S. Röder *et al.*, 2007 A high density barley microsatellite consensus map with 775 SSR loci. Theor Appl Genet 114: 1091-1103.

Volis, S., S. Mendlinger, and N. Orlovsky, 2000 Variability in phenotypic traits in core and peripheral populations of wild barley *Hordeum spontaneum* Koch. Hereditas 133: 235-247.

Volis, S., S. Mendlinger, Y. Turuspekov, and U. Esnazarov, 2002a Phenotypic and allozyme variation in Mediterranean and desert populations of wild barley, *Hordeum spontaneum* Koch. Evolution 56: 1403-1415.

Volis, S., S. Mendlinger, and D. Ward, 2002b Differentiation in populations of *Hordeum spontaneum* Koch along a gradient of environmental productivity and predictability: plasticity in response to water and nutrient stress. Biol J Linn Soc 75: 301-312.

Volis, S., S. Mendlinger, and D. Ward, 2002c Adaptive traits of wild barley plants of Mediterranean and desert origin. Oecologia 133: 131-138.

Volis, S., K. J. Verhoeven, S. Mendlinger, and D. Ward, 2004 Phenotypic selection and regulation of reproduction in different environments in wild barley. J Evol Biol 17: 1121-1131.

von Korff, M., H. Wang, J. Leon, and K. Pillen, 2005 AB-QTL analysis in spring barley. I. Detection of resistance genes against powdery mildew, leaf rust and scald introgressed from wild barley. Theor Appl Genet 111: 583-590.

von Korff, M., H. Wang, J. Leon, and K. Pillen, 2006 AB-QTL analysis in spring barley: II. Detection of favourable exotic alleles for agronomic traits introgressed from wild barley (*H. vulgare* ssp *spontaneum*). Theor Appl Genet 112: 1221-1231.

von Korff, M., H. Wang, J. Leon, and K. Pillen, 2008 AB-QTL analysis in spring barley: III. Identification of exotic alleles for the improvement of malting quality in spring barley (*H. vulgare* ssp *spontaneum*). Mol Breed 21: 81-93.

Walsh, B., 2008 Using molecular markers for detecting domestication, improvement, and adaptation genes. Euphytica 161: 1-17.

Wang, C., S. Zöllner, and N. A. Rosenberg, 2012 A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet 8: e1002886.

Weber, A., R. M. Clark, L. Vaughn, J. de Jesus Sánchez-Gonzalez, J. Yu *et al.*, 2007 Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp. *parviglumis*). Genetics 177: 2349-2359.

Weber, A. L., W. H. Briggs, J. Rucker, B. M. Baltazar, J. de Jesus Sanchez-Gonzalez *et al.*, 2008 The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. Genetics 180: 1221-1232.

Weir, B. S., R. W. Allard, and A. L. Kahler, 1972 Analysis of complex allozyme polymorphisms in a barley population. Genetics 72: 505-523.
Weir, B. S., and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.
White, B. J., C. Cheng, D. Sangare, N. F. Lobo, F. H. Collins *et al.*, 2009 The population genomics of trans-specific inversion polymorphisms in *Anopheles gambiae*. Genetics 183: 275-288.
Wilkes, H. G., 1967 Teosinte: the closest relative of maize. Ph.D. Thesis, Bussey Institute, Harvard University, Cambridge, MA.
Wolfgruber, T. K., A. Sharma, K. L. Schneider, P. S. Albert, D. H. Koo *et al.*, 2009 Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5: e1000743.
Wright, M. H., C. W. Tung, K. Zhao, A. Reynolds, S. R. McCouch *et al.*, 2010 ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. Bioinformatics 26: 2952-2960.
Wright, S. I., and B. Charlesworth, 2004 The HKA test revisited: a maximum likelihood ratio test of the standard neutral model. Genetics 168: 1071-1076.
Wright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley *et al.*, 2005 The effects of artificial selection on the maize genome. Science 308: 1310-1314.
Yang, R. C., 1998 Estimating hierarchical F-statistics. Evolution 950-956.
Yang, W. Y., J. Novembre, E. Eskin, and E. Halperin, 2012 A model-based approach for analysis of spatial structure in genetic data. Nature Genetics 44: 725-731.
Yu, J., J. B. Holland, M. D. McMullen, and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. Genetics 178: 539-551.
Yun, S. J., L. Gyenis, P. M. Hayes, I. Matus, K. P. Smith *et al.*, 2005 Quantitative trait loci for multiple disease resistance in wild barley. Crop Sci 45: 2563-2572.
Yun, S. J., L. Gyenis, E. Bossolini, P. M. Hayes, I. Matus *et al.*, 2006 Validation of quantitative trait loci for multiple disease resistance in barley using advanced backcross lines developed with a wild barley. Crop Sci 46: 1179-1186.
Zohary, D., and M. Hopf, 2000 Domestication of plants in the Old World: the origin and spread of cultivated plants in West Asia, Europe, and the Nile Valley. Oxford University Press, Oxford, UK, 316p